UNIVERSITAT DE BARCELONA

Deep Learning-based Solutions to Improve Diagnosis in Wireless Capsule Endoscopy

Pablo Laiz Treceño

UNIVERSITAT DE BARCELONA

Deep Learning-based Solutions to Improve Diagnosis in Wireless Capsule Endoscopy

Ph.D. THESIS

Pablo Laiz Treceño

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
within the program in
Mathematics and Computer Science

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

July, 2023

Dr. Santi Seguí Mesquida
Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Director

Dr. Jordi Vitrià Marca
Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Co-director

Dr. Santi Seguí Mesquida

Associate Professor of the *Facultat de Matemàtiques i Informàtica* at the *Universitat de Barcelona*, Barcelona.

and

Dr. Jordi Vitrià Marca

Professor of the *Facultat de Matemàtiques i Informàtica* at the *Universitat de Barcelona*, Barcelona.

CERTIFICATES:

That **Pablo Laiz Treceño** has completed his doctoral thesis titled **"Deep Learning-based Solutions to Improve Diagnosis in Wireless Capsule Endoscopy"** under our expert guidance and supervision. The thesis fulfills all the necessary requirements and standards set forth by the *Universitat de Barcelona* for obtaining the Doctoral degree. It is now ready to be presented and defended before the relevant academic panel.

Barcelona, 2023

Dr. Santi Seguí Mesquida

Dr. Jordi Vitrià Marca

To my family,
the ones who have shaped me into the person I am today.

Acknowledgements

I would like to express my deepest and most sincere gratitude to my advisors, Santi Seguí and Jordi Vitrià. From the very beginning, they entrusted me with the opportunity to participate in this remarkable project. Your guidance, support, and encouragement have been invaluable throughout my academic journey.

I am also deeply thankful to all the people who participated in this project, including Hagen Wenzek, Carolina Malagelada, Guillem Pascual, Pere Gilabert, Arnau Quindós, and Reece Moyes. This research would not have been successful without your input and contributions.

I extend my sincere gratitude to my parents, Antonio and Ana, whose constant love and support have been the cornerstone of my personal and academic development. Your encouragement and sacrifices have enabled me to pursue my dreams and achieve this remarkable milestone. I would also like to thank my sister, Miriam, who is the first Ph.D. in our family, for her support, help, and understanding, even when she "may not have received the recognition she deserved". To the rest of my family, who have always been a source of encouragement, I am truly grateful for your support throughout my journey.

My fellow doctoral colleagues, including Guillem Pascual, Pere Gilabert, Paula Gomez, Mariona Caros, Álvaro Parafita, and Axel Brando, have been a constant source of inspiration and camaraderie during these past years. Guillem, it has been a pleasure to share this journey with you. Our conversations, coffees, and laughs have made these years much more enjoyable. Paula and Mariona, thank you for bringing a wonderful dynamic to our group. Without a doubt, these last years with you have been incredible. Pere, thanks for letting us persuade you to join this project. Axel, Álvaro, and Guillem thank you for showing us that completing a Ph.D. successfully is indeed possible. Paula, Pere and Mariona, I have no doubt that you will be next, and I hope to be there to witness your success.

To my closest friends, Noelia and Joan, I cannot thank you enough for being there for me since our teenage years and being a constant presence in my life during the past few years. Your friendship has been a vital source of strength and joy, even during the most challenging times.

I would also like to thank my undergraduate classmates, Marc Beltrán, Sergi Rovira, and Sergi Cebrian. I am grateful for your intellectual stimulation, collaborative spirit,

and shared experiences, which have enriched my experience since my first moment at the University of Barcelona. Additionally, I would like to express my appreciation to Natalia Conesa, Cristian Duran, Laura Portell, and Raquel Colomer, who have laid the foundation of my academic journey. Your friendship, enthusiasm, and dedication to learning have inspired me to be better.

Lastly, I would like to extend my gratitude to my triathlon colleagues, including Albert Segura, Lorena Carreño, Jordi Arguisuelas, and Judith Bueno, for all the shared moments, encouragement, and non-academic teachings that have been of great help in this journey.

Abstract

Deep Learning (DL) models have gained extensive attention due to their remarkable performance in a wide range of real-world applications, particularly in computer vision. This achievement, combined with the increase in available medical records, has made it possible to open up new opportunities for analyzing and interpreting healthcare data. This symbiotic relationship can enhance the diagnostic process by identifying abnormalities, patterns, and trends, resulting in more precise, personalized, and effective healthcare for patients.

Wireless Capsule Endoscopy (WCE) is a non-invasive medical imaging technique used to visualize the entire Gastrointestinal (GI) tract. Up to this moment, physicians meticulously review the captured frames to identify pathologies and diagnose patients. This manual process is time-consuming and prone to errors due to the challenges of interpreting the complex nature of WCE procedures. Thus, it demands a high level of attention, expertise, and experience. To overcome these drawbacks, shorten the screening process, and improve the diagnosis, efficient and accurate DL methods are required.

This thesis proposes DL solutions to the following problems encountered in the analysis of WCE studies: pathology detection, anatomical landmark identification, and Out-of-Distribution (OOD) sample handling. These solutions aim to achieve robust systems that minimize the duration of the video analysis and reduce the number of undetected lesions. Throughout their development, several DL drawbacks have appeared, including small and imbalanced datasets. These limitations have also been addressed, ensuring that they do not hinder the generalization of neural networks, leading to suboptimal performance and overfitting.

To address the previous WCE problems and overcome the DL challenges, the proposed systems adopt various strategies that utilize the power advantage of Triplet Loss (TL) and Self-Supervised Learning (SSL) techniques. Mainly, TL has been used to improve the generalization of the models, while SSL methods have been employed to leverage the unlabeled data to obtain useful representations. The presented methods achieve state-of-the-art results in the aforementioned medical problems and contribute to the ongoing research to improve the diagnostic of WCE studies.

Declaration

I declare that this thesis was composed by myself, that the work contained here is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Pablo Laiz Treceño

Contents

List of Figures

List of Tables

Acronyms

ACC	Accuracy
AE	Autoencoder
AI	Artificial Intelligence
ANG	Angiectasia
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
AV	Ampulla of Vater
BLH	Blood - hematin
BLO	Blood - fresh
CAD	Computer-Aided Detection and Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CAM	Class Activation Map
CCE	Categorical Cross-Entropy
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
CT	Computed Tomography
DL	Deep Learning
DML	Deep Metric Learning
DNN	Deep Neural Network
ERO	Erosion
ERY	Erythema
FB	Foreign body
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive

FPR	False Positive Rate
GAN	Generative Adversarial Network
GI	Gastrointestinal
GIANA	Gastrointestinal Image ANAlysis
GRU	Gated Recurrent Units
IARC	International Agency for Research on Cancer
IBD	Inflammatory Bowel Disease
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoU	Intersection over Union
IV	Ileocecal Valve
KNN	k-Nearest-Neighbors
LES	Lower Esophageal Sphincter
LSTM	Long Short-Term Memory
LYM	Lymphangiectasia
MACC	Mean Accuracy
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MoCo	Momentum Contrast
MRI	Magnetic Resonance Imagery
MSE	Mean Squared Error
MSP	Maximum over Softmax Probabilities
NCE	Noise Contrastive Estimation
NLP	Natural Language Processing
NOR	Normal Clean Mucosa
NT-Xent	Normalized Temperature-Scaled Cross-Entropy Loss
ODIN	Out-of-Distribution Detector for Neural Networks
OE	Outlier Exposure
OOD	Out-of-Distribution
PCA	Principal Component Analysis
POL	Polyp
PYL	Pylorus

RED	Reduced Mucosal View
ReID	Re-identification
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
SENS	Sensitivity
SimCLR	Simple Framework for Contrastive Learning of Visual Representations
SPEC	Specificity
SSL	Self-Supervised Learning
SSP	Sessile Serrated Polyp
SVM	Super-Vector Machines
SwAV	Swapping Assignments Between Views
t-SNE	t-Distributed Stochastic Neighbor Embedding
TL	Triplet Loss
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UES	Upper Esophageal Sphincter
ULC	Ulcer
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder
WCE	Wireless Capsule Endoscopy
WHO	World Health Organization

Chapter 1

Introduction

Contents

The opening chapter of this dissertation emphasizes the motivation behind this research. It aims to grab the reader's attention by presenting convincing arguments for conducting this study. After that, the chapter introduces the main objectives and the important contributions achieved during this doctorate. Finally, there is the outline of this thesis.

1.1 Motivation

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have led to tremendous improvements in medical image analysis, resulting in remarkable enhancements in diagnostic capabilities (Chan et al., 2020b). Both technologies allow the development of methods that possess the ability to learn based on examples (Chan et al., 2020a). These models analyze different input data and search for common patterns to aid in medical diagnosis (Ostherr, 2022).

The potential impact of these innovations on the healthcare sector is profound. By leveraging AI and ML techniques, medical image analysis becomes faster, more accurate, and more accessible (Wang et al., 2021). Hence, these frameworks are expected to revolutionize several aspects of healthcare, such as early disease detection, precise diagnosis, personalized treatment plans, and improved patient outcomes (Chan et al., 2020a; Zhu et al., 2020; Anaya-Isaza et al., 2021; Tufail et al., 2021).

In particular, Deep Learning (DL) techniques, a type of ML algorithm, have emerged as a powerful tool for automatic image analysis and diagnosis, using thousands of previous medical cases and the expertise of hundreds of healthcare professionals (Rana and Bhushan, 2022). The ability to learn from examples enables DL models to extract complex features and make accurate assessments based on medical data (Sejnowski, 2018).

Leveraging the power of DL for medical imaging is not without its challenges. One impediment that arises is the requirement for large numbers of labeled images to train these systems (Saraf et al., 2020). Unfortunately, in the field of medical imaging, datasets are often small and imbalanced. Training models under these circumstances can lead to overfitting and suboptimal performance (Ellis et al., 2022). Moreover, DL models are considered black boxes since humans frequently do not directly interpret their internal workings. This may lead to a lack of confidence from health institutions to adopt the developed technology (Baselli et al., 2020).

These aforementioned problems are typically handled by over-designing systems to overfit the data (Santos et al., 2018). However, it has been demonstrated that such approaches often result in networks that perform poorly with new data as they do not generalize(Varoquaux and Cheplygina, 2022). As a result, alternative strategies are required to overcome these limitations and ensure that DL models can be applied in medical imaging.

This thesis is a compilation of published works that contribute to the ongoing efforts to improve the efficiency and effectiveness of medical diagnosis systems. Particularly, the research is focused on the application of cutting-edge techniques to enhance the performance and capabilities of Wireless Capsule Endoscopy (WCE) (Iddan et al., 2000). WCE is a non-invasive medical imaging technique that has revolutionized the way physicians visualize the Gastrointestinal (GI) tract. It involves swallowing a pill-sized device equipped with a camera that captures an enormous amount of high-resolution images of the digestive tract

as it moves through the body. Manually reviewing these images is time-consuming and often prone to human errors due to the subjectivity involved, making the interpretation process difficult for medical professionals (Beg et al., 2021; Cortegoso Valdivia et al., 2022). The complexity of the digestive system's anatomy presents one of the main challenges owing to the presence of folds, shadows, and reflections in the images. Moreover, the process of distinguishing between normal and anomalous tissue can be also extremely difficult as abnormalities may vary in shape, size, and location (Biniaz et al., 2020).

The vast amount of data generated by WCE requires automatic analysis techniques to perform an efficient and effective interpretation of WCE procedures. CAD systems based on DL are well-suited to achieve this aim. DL models can leverage the large volume of WCE images to develop highly accurate methods capable of identifying abnormalities with high precision (Lan et al., 2019). These systems are expected to significantly reduce clinicians' review time, allowing faster diagnosis while minimizing the risk of misdiagnosis (Lei et al., 2023a). Thereby, DL methods can contribute to better patient outcomes as they enable earlier detection of diseases and more precise treatment planning (Kim and Lim, 2021). Despite the extensive research in this area, further efforts are needed to develop trustworthy and robust systems for clinical settings (Chan et al., 2020a; Muruganantham and Balakrishnan, 2021). In this thesis, several WCE diagnosis challenges are faced to continue with the ongoing improvement of DL-based CAD systems. Particularly, the following problems have been studied: pathology detection, anatomical landmark identification, and Out-of-Distribution (OOD) sample handling.

The first problem addressed is the detection of GI pathologies, with particular emphasis on polyps, which are abnormal tissue growths (Mi et al., 2022). Pathologies and lesions are often small and subtle, making them susceptible to be oversight (Kim et al., 2022). Image artifacts, low image quality, or the presence of similar-looking structures may also lead to misinterpretations (Zhou et al., 2022). Moreover, abnormalities appear sporadically throughout the video, making their detection even more difficult (Bai et al., 2022). Resolving this challenge accelerates the process of reviewing WCE studies and enhances the diagnostic accuracy and efficiency of WCE.

The second challenge approached in this thesis is the identification of anatomical landmarks. They are specific structures on the GI tract used as reference points for medical examinations, surgeries, and other procedures (Lopes et al., 2022). Their proper determination plays a crucial role in the analysis process, since successfully identifying them enables the localization of pathologies and lesions within the digestive system (Koulaouzidis et al., 2021). Automatic detection will speed up the reviewing process of WCE studies.

The third problem lies in building robust and reliable systems. To achieve this aim, two approaches are proposed: the generalization of the model when using data from different devices and the detection of OOD images. In the first case, the model has to work properly even when frames are captured by distinct WCE devices. This meant that the training and test datasets have different distributions caused by the image quality and lighting

conditions, which is referred to as a covariate distribution shift (Guan and Liu, 2021). In the other case, the OOD detection in WCE images aims to discriminate between in-distribution images representing normal anatomy and OOD images showing pathological findings not shown in the training set (Chen et al., 2023). When these problems remain untreated, both scenarios negatively impact the performance of WCE systems analysis, leading to unreliable diagnoses and compromised patient care.

All the challenges faced in WCE have a common thread: the scarcity of labeled data (Vats et al., 2022). The labeling of WCE images requires expert knowledge and meticulous annotation. As a consequence, the size of WCE datasets is usually small, which is a significant obstacle to the design of robust DL algorithms. To overcome this limitation, new approaches have to be developed (Muruganantham and Balakrishnan, 2021).

By tackling these complex challenges, from both DL and WCE, this research aims to improve the performance and applicability of DL-based CAD systems in medical imaging (Altaf et al., 2019; Tai et al., 2023). The outcomes of this work have the potential for enhancing diagnostic accuracy and enable more efficient decision-making processes in healthcare organizations. Moreover, it opens the way for further breakthroughs and improvements in patient care.

1.2 Objectives

The main objective of this thesis is **to develop innovative CAD systems that leverage cutting-edge DL techniques to enhance the diagnosis of WCE studies by reducing the required review time while preserving or increasing their performance**.

The specific goals of this thesis are summarized as follows:

1. **Create models to tackle data scarcity**. This involves developing different approaches that can learn and extract better features than standard models, even when there is a small amount of labeled data to train a neural network.

2. **Create a model to tackle imbalanced datasets**. This entails creating DL methods capable of handling imbalanced datasets, that is, where some classes have significantly fewer examples than others do.

3. **Create a model that integrates explainability**. This implies the development of CAD systems that provide methods to understand the logic behind their predictions.

4. **Create a model to detect GI pathologies**. This involves developing CAD systems to detect and locate pathologies, with special attention to polyps in images acquired from WCE videos.

5. **Create a model to identify anatomical landmarks**. This involves creating a clinical decision support tool that can identify anatomical landmarks in images acquired from WCE videos.

6. **Create a model that generalizes with data from different devices**. This entails creating a reliable CAD system capable of effectively analyzing images captured from a variety of WCE devices, regardless of the manufacturer or version. It should maintain optimal performance and be able to handle any differences in image characteristics.

7. **Create a robust model to detect OOD samples**. This involves creating a DL model capable of detecting and flagging findings that are not part of its training data.

To confirm that each method contributes significantly to the field of medical imaging and helps to improve the diagnosis of GI conditions, it must fulfill the following requirements and considerations:

1. **State-of-the-art results.** The DL models will achieve state-of-the-art results in terms of relevant metrics for the specific task. These results should be compared with other approaches or existing systems to demonstrate their superiority and potential benefits.

2. **Limitations.** The study of the method will include clear identification and discussion of its limitations. In addition, it should contain how these limitations will be addressed in future iterations of the system.

1.3 Contributions

This section compiles the contributions to the field of WCE diagnosis carried out in this thesis.

- Polyp detection in WCE images is a challenging task due to a variety of factors, including the diverse appearance of polyps, imbalanced datasets, and limited data availability. To address this issue, a DL model optimized with the Triplet Loss (TL) is used to enhance the feature extraction capability. The experimental results show that the proposed system outperforms existing state-of-the-art methods. Furthermore, an explainability technique is implemented to provide insights into the decision-making process of the model. This method was published in the prestigious journal *Computerized Medical Imaging and Graphics* (Laiz et al., 2020).

- Problems related to WCE datasets are notoriously challenging due to the limited number of labeled samples and imbalanced classes, which present significant obstacles in developing accurate models. To tackle these matters, this study introduces a novel approach that combines Self-Supervised Learning (SSL) with pseudo-label extraction from unlabeled videos. The initial parameters are optimized by utilizing a pretraining phase with SSL. The compressed representations generated by the model contain valuable information that can be used for multiple downstream tasks. In particular, this knowledge is used in two cases: polyp detection and inflammatory and vascular lesion identification. In both tasks, state-of-the-art results are achieved. This research was published in the reputable journal *Computers in Biology and Medicine* (Pascual et al., 2022a).

- Identifying anatomical landmarks accurately is crucial for effectively interpreting WCE videos. For this reason, a novel system is developed in this study. It combines images, timestamps, and motion data to precisely detect the capsule's entrance and exit in both the small bowel and large intestine. The proposed method is evaluated on three distinct datasets, and the results show significant improvements over the state-of-the-art systems. The research paper was accepted for publication in the journal *Computerized Medical Imaging and Graphics* (Laiz et al., 2023).

- The mitigation of the covariate distribution shift is important to avoid the deterioration of the performance of DL models. The following research explores the use of the TL as a domain adaptation technique to leverage frames from both training and test datasets, that are extracted from different WCE devices, respectively. The approach aims to improve the overall performance and the model's generalization. The experimental results show that the proposed domain adaptation approach outperforms transfer learning techniques and training the system from scratch. The resulting method was presented in the *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (Laiz et al., 2019).

- OOD image detection allows to design robust and reliable DL systems. To achieve this, a novel OOD detector for WCE images is proposed. The method is a patch-based methodology that involves three stages: (1) training a self-supervised triplet network to learn representations of WCE images using patches; (2) clustering the patch embeddings based on visual similarity; and (3) using the cluster assignments as pseudo-labels to train a patch classifier and detect OOD images using ODIN. The proposed method is evaluated with the public dataset *Kvasir-Capsule* (Smedsrud et al., 2021). The results show that the method achieves higher performance than the tested methods. This research was published in the reputable journal *Artificial Intelligence in Medicine* (Quindós et al., 2023).

1.4 Outline

This thesis is organized as follows. The next chapter provides the background of this dissertation. It starts with an outline of the digestive system, with a particular emphasis on lesions and pathologies of the GI tract. After that, the WCE is presented along with its reading process and the technique's limitations. The subsequent section provides a general overview of DL and its applications in medical imaging. Finally, there is a brief introduction to CAD systems in WCE.

A comprehensive overview of the theoretical foundations of DL and tools used to achieve the set goals is presented in Chapter 3. It is divided into four sections: Deep Metric Learning (DML) and contrastive approaches, SSL, OOD samples, and a summary of the key takeaways.

As the thesis is a compilation of publications carried out during this doctorate, the following chapters contain different solutions to three diagnostic problems in WCE.

The first problem focuses on the detection of GI pathologies, mainly colorectal polyps. Polyp detection is a complex task owing to the range of positions, sizes, forms, and locations. However, DL-based CAD methods are well-suited to accomplish a reliable identification based on their visual features. Moreover, the lack of annotated data, the imbalanced datasets, and the requirement for model interpretability must also be addressed. Chapters 4 and 5 present two papers that propose DL methods to achieve this aim.

The second explored problem is the identification of anatomical landmarks, which are crucial points of reference for physicians in diagnosing and managing various GI disorders. The complexity of this task relies on identifying one frame for each landmark. In particular, Chapter 6 focuses on the anatomical landmark detection of the small intestine and colon.

The last two publications, displayed in Chapters 7 and 8, provide solutions to two problems related to OOD samples. Chapter 7 presents a methodology that achieves the generalization of the system, even when test images come from a different source (device) than training data. In contrast, Chapter 8 proposes an OOD detector, which seeks to discover GI pathologies that were not included in the training data.

To end the dissertation, the overall results of the developed studies are discussed in Chapter 9. Finally, in Chapter 10, the conclusions and suggestions for further research are presented.

Chapter 2

Background and Related Work

Contents

To establish the background and provide the reader with the necessary knowledge to understand the studied problems, the chapter is organized as follows. First, an overview of the digestive system is provided, along with the most important lesions and pathologies of the GI tract, as well as the anatomical landmarks of its organs. After that, WCE and the reading process of this technique along with their limitations are presented. Then, there is a general introduction to DL and medical imaging, including the difficulties encountered in this field. Finally, there is a brief discussion of the related work of CAD systems in WCE.

2.1 Digestive System

The digestive system is a set of organs that participate in food intake, digestion, and nutrient absorption. The organs ensure that nutrients extracted from meals reach the cells of the body to maintain the immune system and assist in other vital functions. It also eliminates the food and product waste from numerous endogenous metabolic processes (Smith and Morton, 2010).

The anatomy of the digestive system, illustrated in Figure 2.1, can be split into two groups. The mouth, esophagus, stomach, small intestine (small bowel), and large intestine (colon) compose the initial group of organs known as the primary organs. The second group is known as the accessory organs of digestion and includes the liver, pancreas, and gallbladder.

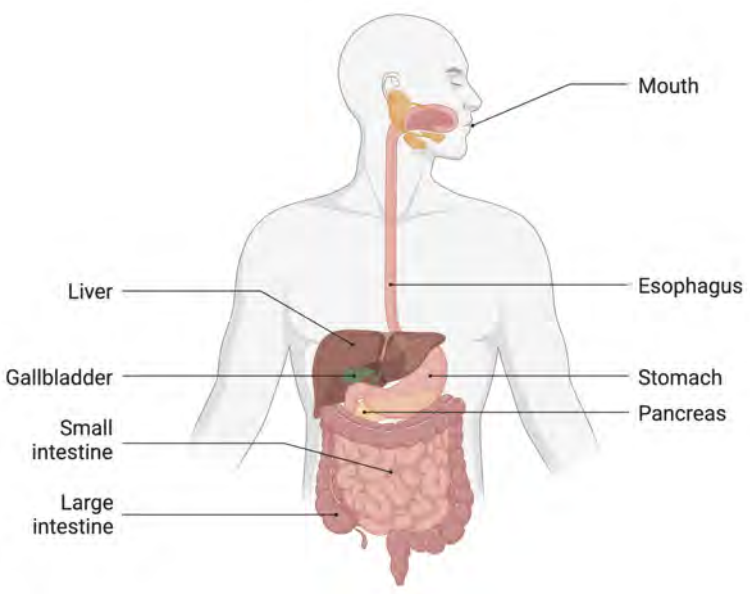
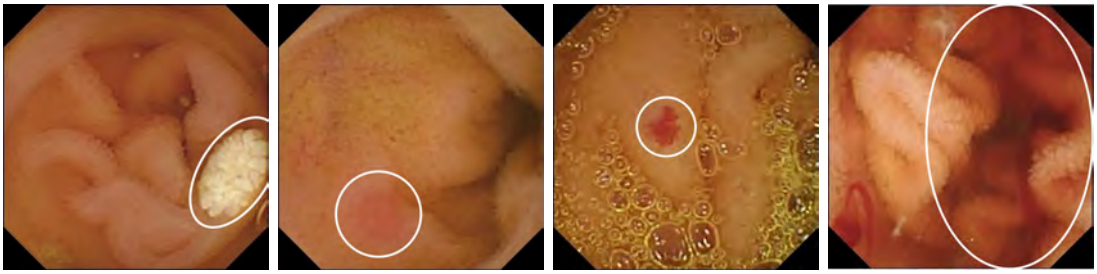


Figure 2.1: Anatomy of the digestive system. Image adapted from *BioRender*.

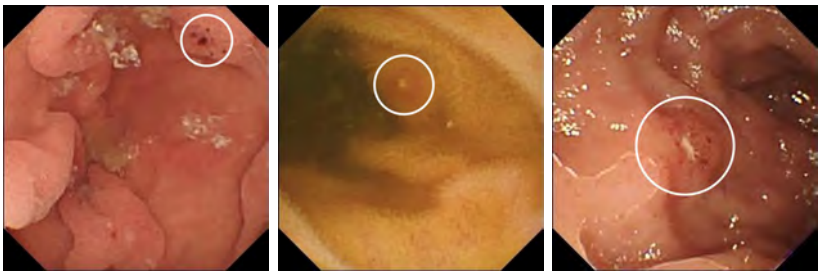
The person's health and well-being are closely related to a well-functioning digestive system, which occurs when there is a balanced diet, regular physical activity, and stress management. When this is not the case, lesions and disorders can emerge. Several digestive screening techniques are available for assessing the digestive system's health. Depending on the patient's needs, diagnostic tools range from simple techniques like stool tests to more complex methodologies like medical imaging. Endoscopy and colonoscopy are the gold-standard screening procedures when a visual examination of the digestive tract is required. They allow the visualization of the upper GI tract or the large intestine, respectively. During the procedure and depending on the location, a flexible tube ending in a camera is inserted through the mouth or rectum to screen for GI disorders.

2.1.1 Digestive System Pathologies

The digestive system can be affected by many different GI disorders, from common conditions such as acid reflux or Inflammatory Bowel Disease (IBD) to more serious diseases such as bleeding or Colorectal Cancer (CRC). Although the early detection of polyps is one of the main focuses of this thesis, other lesions and pathologies such as lymphangiectasia, erythematous mucosa, angiectasias, GI bleeding, erosions, and ulcers are also studied. Figure 2.2 and Figure 2.3 show visual examples of each of them.



(a) Lymphangiectasia (b) Erythematous mucosa (c) Angiectasia (d) Fresh blood



(e) Hematin blood (f) Erosion (g) Ulcer

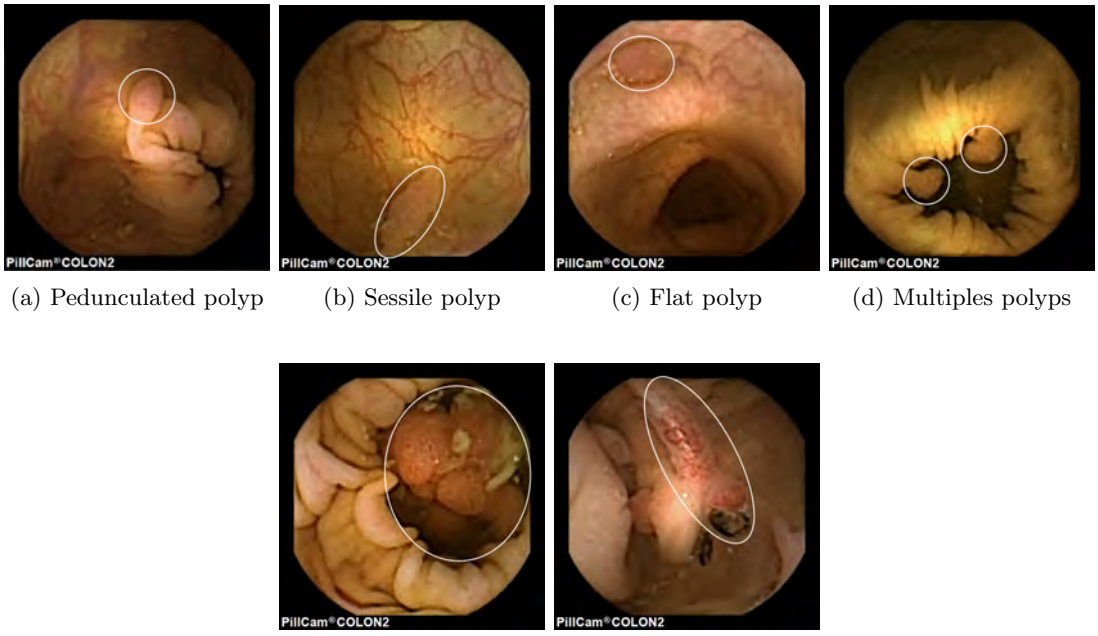
Figure 2.2: Mosaic of images with examples of the different pathologies explained in Section 2.1.1. Images extracted from the dataset of Smedsrud et al. (2021).

Lymphangiectasia (Figure 2.2a) is a condition characterized by dilated lymphoid vessels in the mucosal wall. Erythematous mucosa (Figure 2.2b) corresponds to a reddish appearance of the mucosa, whereas angiectasias (Figure 2.2c) are small, superficial, dilated vessels that can cause chronic bleeding and anemia, particularly in persons with chronic heart and lung diseases.

Lesions in the upper GI tract or small bowel can produce GI bleeding, resulting in the appearance of fresh, red-colored blood (Figure 2.2d). In cases of minor bleeding, black stripes of blood called hematin (Figure 2.2e) can be observed on the mucosa surface. Excavated lesions in the mucosa can cause erosion (Figure 2.2f), which may be covered by a tiny fibrin layer or may extend to form larger ulcers (Figure 2.2g).

Polyps and Colorectal Cancer

A colonic or colorectal polyp is the precursor lesion of CRC. It is a small abnormal growth of tissue that protrudes from the surface of the colon's mucosa membrane, most commonly in the sigmoid colon and rectum. These growths can be either benign (non-cancerous) or malignant (cancerous), and their size and shape can vary widely, as seen in Figures 2.3a–2.3d. Moreover, they can be found as single, as shown in Figure 2.3a, or multiple, as shown in Figure 2.3d.



(a) Pedunculated polyp (b) Sessile polyp (c) Flat polyp (d) Multiples polyps

(e) CRC (f) CRC

Figure 2.3: Mosaic of images with examples of four different polyps (2.3a - 2.3d) and two CRC images (2.3e - 2.3f).

Polyps are generally classified according to their visual characteristics, with size and morphology being the most relevant. Based on their diameter, they are categorized as diminutive (less than 6 mm), small (6 mm to 9 mm), or large (more than 10 mm) (M. Meseeha, 2022). Polyps, on the other hand, are also divided into depressed, flat, sessile, or pedunculated depending on their morphology (Inoue et al., 2003). Each case is illustrated in Figure 2.4.

Polyps are diagnosed through screening techniques such as colonoscopy or WCE. Their visual features and location play a crucial role in assessing their type and potential to become malignant. However, focusing only on their aspect is insufficient for accurately identifying their nature and danger. Hence, a biopsy is usually required to confirm the diagnosis. When a polyp is detected, physicians surgically remove it and dispatch it to a laboratory for pathological analysis. If the lesion is benign, no further treatment is

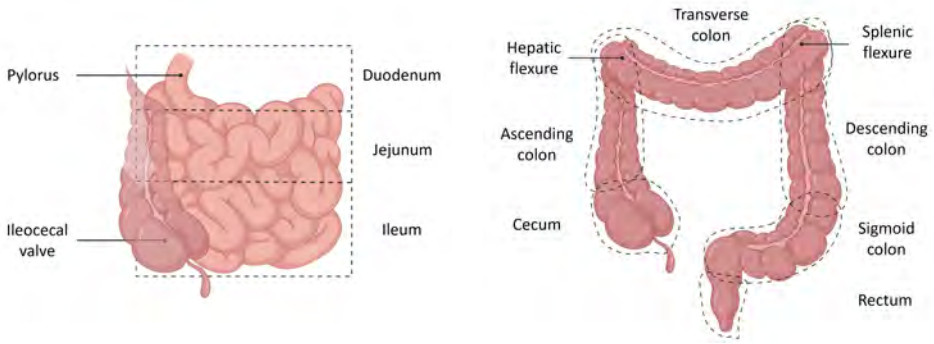
(a) Depressed (b) Flat (c) Sessile (d) Pedunculated

Figure 2.4: Morphology of polyps. Images adapted from *United European Gastroenterology*.

required. Otherwise, if the polyp is malignant or has the potential to be, treatment options may include surgery, chemotherapy, or radiation therapy. In cases where the polyp is left untreated, it can develop into CRC, the third most common cancer in the world (World Health Organization, 2022). Examples of it are shown in Figure 2.3e and 2.3f. Therefore, regular screening and early detection are critical to prevent further complications.

2.1.2 Anatomical Landmarks

The anatomical landmarks of the digestive system are specific structures that subdivide the GI tract. They are used as reference points for identifying locations and determining the relationship between different organs. These structures aid physicians in the diagnosis and treatment of certain conditions through their localization. However, these landmarks are difficult to accurately identify as they are approximate regions with similar-looking features.



(a) Small bowel (b) Large intestine

Figure 2.5: Anatomy of the small bowel (a) and the large intestine (b). The regions delimited by dashed lines are the parts of the organs, whereas the lines ending with a dot refer to the anatomical landmarks. Images adapted from *BioRender*.

The small bowel anatomy is divided into three parts: the duodenum, the jejunum, and the ileum, approximately as depicted in Figure 2.5a. The duodenum is the small intestine entrance. The jejunum is the middle part of the organ. Finally, the ileum is located at the end of the small bowel.

The pylorus and the ileocecal valve, which limit the organ, are two key landmarks

for gastroenterologists. The pylorus is the opening at the lower end of the stomach that connects to the small intestine. The ileocecal valve is a ring-shaped muscle that connects the ileum and the cecum (the first part of the large intestine).

On the other hand, the large intestine is divided into the cecum, ascending colon, transverse colon, descending colon, sigmoid colon, and rectum, as illustrated in Figure 2.5b. The cecum is the first part of the organ and has a pouch-like structure. The ascending colon is situated next to it. The transverse colon, the colon's longest segment, is located across the abdomen. The descending colon connects the transverse and the sigmoid colon. The latter is an S-shaped region of the large intestine that ends in the rectum.

The landmarks of the large intestine are the hepatic and splenic flexures. The first is the curve where the ascending colon becomes the transverse colon. The second is the bend where the transverse colon becomes the descending colon.

2.2 Wireless Capsule Endoscopy

Endoscopy and colonoscopy are the gold-standard screening procedures to examine the digestive system. However, both procedures carry some associated risks and issues (Häfner, 2007; Sieg, 2011). Both techniques are invasive and might cause discomfort and pain in patients, requiring sedation in some cases for management. Anesthesia, on the other hand, can result in unpleasant side effects such as nausea, drowsiness, and confusion. Complications, such as hemorrhage, perforation, and infection, are also possible with these techniques. Moreover, GI preparation is required before the procedure, which can cause discomfort and dehydration. In addition to physical discomfort, both procedures demand a considerable amount of time and resources for preparation and execution.

Wireless Capsule Endoscopy (WCE) (Iddan et al., 2000) is a relatively new technology, approved by the Food and Drug Administration (FDA) in 2001 (Food and Drug Administration, 2001). It is a non-invasive diagnostic technology used to visualize the inside of the GI tract. The small, pill-sized capsule contains a camera, a light, and a transmitter. The patient swallows the capsule, like the one shown in Figure 2.6, which travels through the GI tract and captures images of the interior of the digestive system. There is a small portable recording device known as a data recorder that receives the signals transmitted by the capsule. This device is attached to the patient's belt during the examination, allowing him to continue with his daily life while the analysis is being carried out. Once the procedure is complete, a workstation with pre-installed software is available to physicians, allowing them to view and interpret the images captured by the capsule.

Figure 2.6: PillCam SB3 Capsule. Image from *Medtronic*.

At the start of the 21st century, the medical community welcomed WCE with great enthusiasm, as it expanded the diagnostic capabilities in the digestive system, particularly in the small intestine. Compared with endoscopy and colonoscopy, WCE is less intrusive and causes less discomfort to patients. Additionally, it provides a complete visualization of the entire GI tract, including hard-to-reach areas that endoscopy and colonoscopy may not be able to access. As a result, WCE has become a significant tool for diagnosing. In particular, it is commonly used in patients experiencing symptoms such as abdominal pain, GI bleeding, or anemia without a known cause.

Currently, small bowel WCE is used as the initial diagnostic tool in several countries, including the United States, the United Kingdom, and across various European nations for cases of GI bleeding (Goyal et al., 2022; Garbaz et al., 2022), Crohn's disease (Xing and Mouchère, 2022; Saxena et al., 2022), and, to a lesser extent, polyposis syndromes

(Falin et al., 2022), celiac disease (Sana et al., 2020), or other small bowel pathologies (Trasolini and Byrne, 2021). On the other hand, colon WCE is gaining acceptance as a reliable method for detecting polyps (Pascual et al., 2022b; Reuss et al., 2022; Gilabert et al., 2022), bleeding (Shahril et al., 2020; Smedsrud et al., 2021; Rustam et al., 2021), and ulcers (Wang et al., 2019b; Aoki et al., 2019; Wang et al., 2019c), as well as completing an unfinished colonoscopy (Yung et al., 2016; Koulaouzidis et al., 2021).

The initial commercial version of WCE, PillCam SB, was a compact device measuring 26 mm in length, 11 mm in diameter, and weighing less than four grams. The capsule captured two frames per second, translating to $57,600$ frames after 8 hours of analysis. Since then, numerous manufacturers have released new devices with different specifications, such as size, weight, frame rate acquisition, battery life, camera resolution, and reading software. Although most capsules are designed for small bowel and colon examination, Given Imaging has also produced capsules specifically for analyzing the esophagus. Small-bowel capsules typically include a single camera. However, devices that focus on the colon contain two cameras, one at each end, to provide a complete view of the colon's interior. The company CapsoVision, on the other hand, redesigned the capsule to include four cameras in the middle, providing a 360-degree field of view. Further details on the different capsules and their specifications are summarized in Table 2.1 (Kurniawan and Keuchel, 2014).

Table 2.1: Summary of commercial WCE devices used to perform screening procedures. Data extracted from Ciuti et al. (2011); Wang and Meng (2011); Kurniawan and Keuchel (2014); CapsoVision (2022). The legend used is: Length (L) and Diameter (D).

Capsule Name	Manufactures	Year	Size (L×D) (mm)	Weight (g)	Battery life (h)	Organ	Frames per second	Field of view (^{o})
PillCam SB	Given Imaging	2001	26.0×11.0	4.00	8	Small Bowel	2	156
PillCam SB2	Given Imaging	2012	26.2×11.4	3.40	8	Small Bowel	$2-6$	156
PillCam SB3	Given Imaging	2015	26.0×11.0	3.00	>8	Small Bowel	$2-6$	156
PillCam COLON	Given Imaging	2006	31.4×11.4	2.90	10	Colon	4	154
PillCam COLON 2	Given Imaging	2012	31.5×11.6	2.90	10	Colon	$4-35$	172
PillCam CROHN	Given Imaging	2018	26.0×11.0	2.90	>10	Intestines	$4-35$	172
PillCam ESO	Given Imaging	2004	26.0×11.0	4.00	1/2	Esophagus	14	140
PillCam ESO 2	Given Imaging	2008	26.0×11.0	4.00	1/2	Esophagus	18	169
EndoCapsule	Olympus America	2012	26.0×11.0	3.50	>8	Small Bowel	2	145
MiroCam	IntroMedic Company	2014	24.5×10.8	4.70	>11	Small Bowel	3	170
OMOM Jinshan	Science and Technology	2005	27.9×13.0	6.00	>8	Small Bowel	2	140
CapsoCam	CapsoVision	2016	31.0×11.0	4.00	15	Small Bowel	20	360

The unique characteristics of each capsule result in visual variations in the captured frames. New devices obtain images with better clarity and sharpness, more color information due to the lighting, or less blurring. An example of the images captured by each capsule is shown in Figure 2.7. In particular, Given Imaging capsules (Figures 2.7a–2.7g) contain similar lenses and, as a result, a comparable profile. In contrast, the shape of EndoCapsule images (Figure 2.7h) is similar to the one obtained during colonoscopies. Standing out from the others, the CapsoCam (Figure 2.7k) images show a 360-degree field of view.

(a) PillCam SB (b) PillCam SB2 (c) PillCam SB3 (d) PillCam COLON (e) PillCam COLON 2

(f) PillCam CROHN (g) PillCam ESO 2 (h) EndoCapsule (i) MiroCam (j) OMOM Jinshan

(k) CapsoCam

Figure 2.7: Mosaic with random captured frames of the different capsules endoscopy devices introduced in Table 2.1.

2.2.1 Reading Process

A WCE reading process involves at least one gastroenterologist reviewing and interpreting the large volume of images captured by the capsule as it moves through the GI tract. The interpretation of the thousands of generated images is a time-consuming and labor-intensive task for endoscopists. Moreover, the reviewing process of a WCE video requires multiple stages to guarantee that all relevant information is analyzed. It is fair to say that this reading process might be repetitive, tedious, and even monotonous, especially when the video is long or negative.

Currently, there is a lack of scientific evidence for determining the optimal protocol for reading a WCE video (Adler et al., 2015; Yamamoto et al., 2017; Enns et al., 2017; Rondonotti et al., 2020; Koulaouzidis et al., 2021). Hence, no standards or guidelines have been established. Until now, there has been only a combination of the suggestions provided by capsule manufacturers and the opinions of experts in the field (Adler et al., 2015; Yamamoto et al., 2017; Ahmed, 2022). These recommendations can be summarized in three steps: the patient's clinical history, a quick overview of the video, and a review.

Clinical History

Before reading any WCE video, it is essential to be aware of the patient's clinical information (Rondonotti et al., 2020), such as clinical presentation details, duration of the clinical history, existence of one or more additional medical conditions, medications, and, most importantly, the referrer's query. The knowledge of the patient's medical history and symptoms can help guide the interpretation of the images, which is crucial for a balanced evaluation of findings and meaningful WCE conclusions (Koulaouzidis et al., 2021).

Quick overview of the video

After obtaining the patient's clinical history, it is recommended that physicians conduct a quick overview of the entire video. The goal of this initial phase is to quickly identify areas of interest for a more in-depth study, locate key anatomical landmarks, and obtain vital clinical information in a timely manner. This can be done using automated fast-reading modes that are specifically designed for this purpose (Saurin et al., 2012; Hausmann et al., 2019; Freitas et al., 2020).

The landmarks that should be located in the small intestine are the pylorus and ileocecal valve to assess small bowel cleansing, confirm complete small bowel exploration, and precisely point potential lesions (Rondonotti et al., 2020). In the colon, the cecum and rectum, which delimit the organ, must be indicated. Furthermore, the hepatic and splenic flexures are important landmarks to place, as this information is used to assess screening completion (Koulaouzidis et al., 2021). The relative position of the small bowel and the colon landmarks within the video can also be used to adjust the frame speed per segment, accelerating the diagnostic process (Koulaouzidis et al., 2021). However, it is important to note that accurate and proper landmarking can be a challenging task due to the to-and-fro movement of the capsule as well as inadequate bowel preparation in some patients.

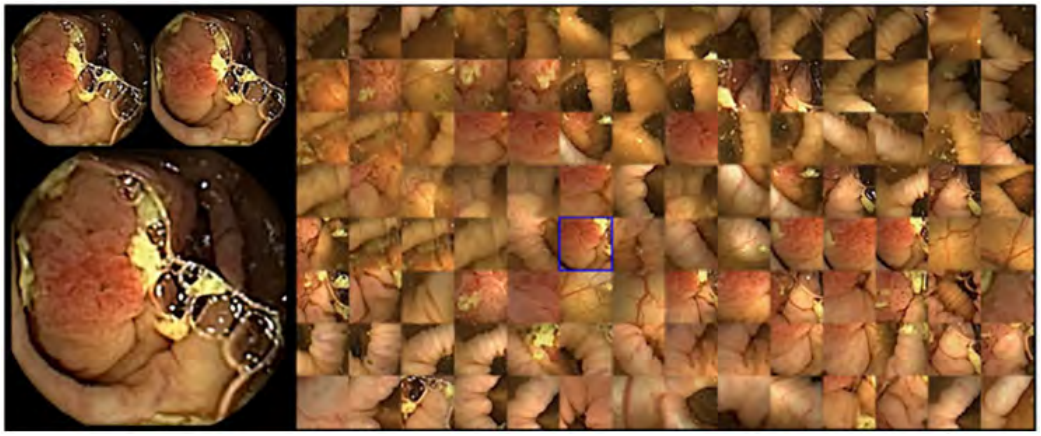
Review

After the first quick overview, the goal of the review is that experts mark any suspected lesions and take some normal images for photo-documentation to aid in the diagnosing procedure. To ensure that all relevant information is captured, it is important to annotate findings using dedicated classification or scoring systems. Additionally, it is necessary to register the overall adequacy of bowel cleanliness in the report, because inadequate visualization can increase the missed rate of lesions (Spada et al., 2021).

Although a complete protocol for the review process does not exist, there are some recommendations depending on the organ studied. For the small intestine, it is suggested that readers avoid using high frame rates (Rondonotti et al., 2020), as this can make identifying minimal changes difficult, even for expert readers. In the colon, it is advised to review the images from one camera at a time, followed by the other. The suggested rate is $8 - 15$

frames per second, so it can be slowed down if the capsule is moving too fast. In both cases, it is recommended to use different AI software readers to check for missing findings (Koulaouzidis et al., 2021). Examples of these systems are the *Top 100* mode (Freitas et al., 2020), which shows candidate frames containing small polyps, and the *Collage* mode, which selects and displays suspicious frames from the video. An example of both modes using the *RAPID* Reader software of Given Imaging is shown in Figure 2.8.

(a) *Top 100* mode. Using AI, 100 candidate frames to contain small polyps are shown.



(b) *Collage* mode. Using AI, suspicious frames are selected and shown.

Figure 2.8: Example of two reading modes: Top 100 and Collage Mode from *RAPID* Reader software of Given Imaging. Images extracted from Koulaouzidis et al. (2021).

The GI has propulsive activity; hence, it is common for the capsule to move in both directions, which means that the same lesion or anomaly may appear several times. Consequently, the second part of the review should focus on comparing the saved findings and unifying them in unique lesions. As an optional step, a final confirmatory run on the annotated findings should be performed in reverse mode, that is, from the rectum to the top, to confirm the findings and eliminate duplicates.

Before writing the report, physicians should systematically examine the other organs

to identify any possible lesions. Between 7% and 15% (up to 30% in certain reports) of the time, patients have lesions outside the studied organ (Koffas et al., 2018). It is also suggested to have a "think twice" policy in negative exams, where the same reviewer or, if possible, a second, more experienced one repeats the review process.

Pre-readers and readers

As can be concluded from the previous details, the reading process of WCE videos is a complex task, especially considering that relevant information appears in a few frames of the entire screening. For this reason, despite the scarcity of studies, the use of a reader-validation team is highly recommended during this process (Barkin and Barkin, 2017; Rondonotti et al., 2020), particularly in colon screening. This approach employs one reviewer as a pre-reader and another as a validator. (Koulaouzidis et al., 2021). The WCE recording is reviewed by the pre-reader, who is usually an experienced nurse, who adds landmarks and any obvious pathology, which is then validated by the reader. It is advised that pre-readers also highlight areas of ambiguity where the validator's opinion is required.

2.2.2 Drawbacks

WCE is a valuable tool for the diagnosis of GI disorders and has gained popularity in the medical community, but it is not without its drawbacks. From a design perspective view, the capsule may not always be entirely conclusive due to factors such as intestinal content or a limited field of view. Furthermore, it does not execute therapeutic procedures, meaning that if further tests are needed, additional techniques must be employed (El-Matary, 2008).

For physicians, the interpretation of thousands of WCE images is an overwhelming and time-consuming task, which can lead to fatigue (Lan et al., 2019; Beg et al., 2021). The danger of fatigue involves an increase in the likelihood of missing clinically relevant findings that appear in only a few frames. Indeed, Lewis et al. (2005) reported that the rate of missed lesions in these cases is around the 10%.

Furthermore, the organ where the capsule is located at a certain time is unknown. Consequently, the exact location of the visualized lesion is difficult to discern. Anatomical landmarks are commonly employed to limit organs and thus, allow the estimation of the location of the pathologies. Moreover, they can be used to adjust the frame rate during the reading process (Koulaouzidis et al., 2021). However, the accurate identification of anatomical landmarks is a complex task due to the many similar-looking structures of the digestive tract and the limited visual field of the capsule camera, in a constantly moving environment. Moreover, the view can also be obstructed by the mucosal folds lining the organs, making it even more challenging to locate them properly.

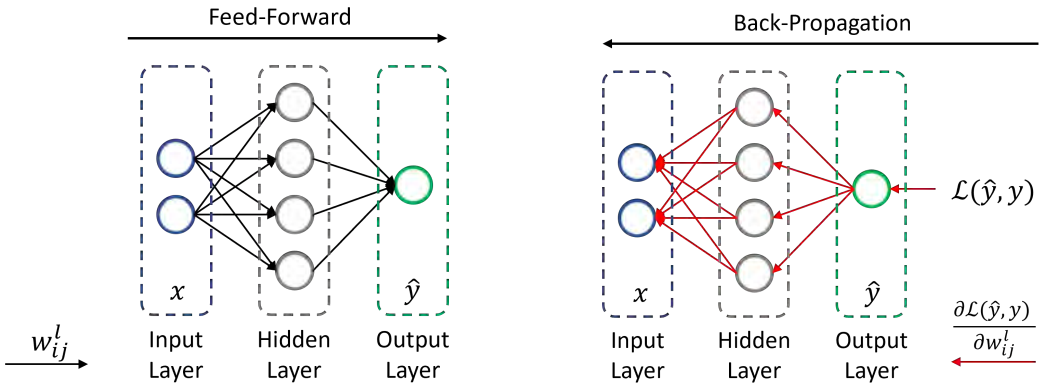
To address the two last drawbacks and improve the clinical utility of WCE, various solutions, including the use of AI and DL techniques, are being explored. To overcome

the large volume of images that have to be reviewed by physicians, new technologies and techniques are being studied to accelerate the image review process and reduce the risk of missed lesions. In particular, DL approaches are well suited to this aim, as they can automatically learn to recognize patterns in the data. These techniques can aid in object detection and image classification, assisting physicians in identifying relevant findings more quickly (Gilabert et al., 2022). Regarding the location of pathologies in the GI tract, various techniques are being developed to pinpoint, or at least limit, the area of the lesions. As mentioned previously, it can be accomplished by locating anatomical landmarks. It is achieved by using software-assisted detection algorithms that aid in the classification of images into organs and then in the identification of landmarks. Overall, while WCE has its limitations, ongoing advancements in DL hold promise for overcoming the drawbacks, leading to improved clinical outcomes and patient care.

2.3 Deep Learning and Medical Imaging

Deep Learning (DL) is a sub-field of ML that aims to learn the underlying patterns in the data, allowing the systems to make sophisticated decisions (LeCun et al., 2015; Goodfellow et al., 2016). The concept of DL was first introduced in the 1980s; however, it only began to gain mainstream attention in the early 2010s with the advent of powerful computational resources and extensive data collection. The ability of DL models to process and learn from data has led to significant breakthroughs in a wide range of fields, including computer vision (Voulodimos et al., 2018), Natural Language Processing (NLP) (Qiu et al., 2020), and decision-making (Latif et al., 2019).

DL involves the use of models called Artificial Neural Networks (ANNs) (McCulloch and Pitts, 1943). They are inspired by the structure and function of the brain and are able to analyze and process large amounts of data. They are formed by interconnected nodes, named artificial neurons. Each pair has an associated weight w_{ij}^{l}, which is adjusted to perform a specific task. Neurons are organized into layers and are typically divided into three main categories: input, hidden, and output. An example of it is illustrated in Figure 2.9 where the blue dotted square represents the input layer, the hidden is the grey one which must be between the other two, and finally, the output, displayed as a green dashed square. The latter is responsible for providing the predictions of the system. It is worth noting that when the number of hidden layers is large in an ANN, the network is referred to as Deep Neural Network (DNN).



(a) Feed-Forward (b) Back-Propagation

Figure 2.9: Principle of training an ANN. In (a) the arrows between the neurons represent the weight of the ANN, w_{ij}^{l}, used to infer the network output. In (b) the arrows represent the gradient of the loss with respect to the weights. Using both steps, the weights are adjusted to obtain a more accurate network prediction.

The learning process of an ANN is based on the idea of adjusting the network's weights to improve the accuracy of the predictions. To this aim, two steps are required. First, the input layer of the ANN receives the raw data, which is then processed and transformed

by the hidden layer(s) before being transmitted to the output layer for a final prediction or decision. This process is called feed-forward or inference, as illustrated in Figure 2.9a. In the second step, the adjustment of the weights is done using a loss function and the back-propagation algorithm (Rumelhart et al., 1986). The loss function computes the error between the network's prediction and the ground truth, normally designated as $\widehat{y}$ and y, respectively. Different loss functions can be applied depending on the problem, data, and neural network used. However, the most common losses in computer vision problems are Categorical Cross-Entropy (CCE), Mean Absolute Error (MAE), and Intersection over Union (IoU). The back-propagation algorithm is an optimization technique that enables ANN to learn from data. It minimizes the error produced by the network, that is, the value obtained after applying the loss function. This algorithm propagates the errors to the input data and updates the parameters using the gradient of the loss with respect to the weights. A schematic illustration of the process is shown in Figure 2.9b. The two-step process is repeated until the neural network is able to accurately predict the output for the given input data.

DL techniques have been applied to a wide range of problems, reaching state-of-the-art results in many areas. Particularly, notable success has been accomplished in the field of computer vision, where Convolutional Neural Networks (CNNs) (Fukushima, 1980; Lecun et al., 1998) have been used to achieve near-human performance in tasks such as image classification (Yadav and Jadhav, 2019), and object recognition (Xiao et al., 2020). They are inspired by the organization of the visual cortex and the way it processes information from the eyes. As a consequence, they are designed to process images and preserve the spatial structure of the data. These networks consist of a combination of different layers, including convolution, pooling, and dense. The convolutional layers have sets of 2D filters, also called kernels. The weights of the filters are optimized with the back-propagation algorithm to identify patterns and characteristics in the images like edges, corners, and textures by using the mathematical convolution operation. Pooling layers reduce the input's spatial dimensions by summarizing local features and preserving the most relevant information. And finally, dense layers connect each neuron between the previous and the current layers.

Since its first appearance, many variants of CNNs have been published, and the most outstanding ones are presented below. Lecun et al. (1998) introduced LeNet-5, one of the first CNN designed for handwritten digit recognition. In 2012, Krizhevsky et al. (2012) presented AlexNet, which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) that year. The authors introduced three additional convolutional layers to capture more complex features than LeNet-5. Moreover, the use of max-pooling layers to downsample the feature maps (Scherer et al., 2010), and the adoption of the Rectified Linear Unit (ReLU) activation function (Nair and Hinton, 2010), led to a significant improvement in the performance. Later, VGGNet was introduced by Simonyan and Zisserman (2015) and the main difference with AlexNet was the concatenation of multiple convolutional layers before applying a max-pooling operation to achieve a deeper architecture than before. Because of its structure, VGGNet is widely used as a feature extractor in segmentation and

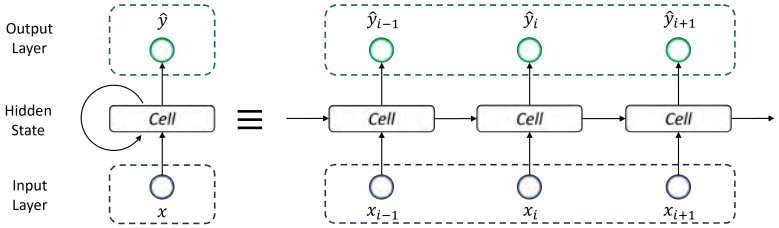
object detection tasks (Geng et al., 2019; Haque et al., 2019). In the same year, Szegedy et al. (2015) proposed GoogleNet, known for its use of inception modules. The modules are designed to generate a single output by concatenating the results from several convolution operations with multiple filter sizes in parallel. This approach allows the network to learn multiple feature scales simultaneously.

He et al. (2016) proposed a DNN called ResNet, which also won the ILSVRC. The main innovation made by the authors was the use of residual connections, which add the input of n consecutive convolutional layers to its output, enabling the layer to learn a residual mapping between the input and output. This strategy provides networks with more layers that converge without suffering from exploding or vanishing gradient problems (Grosse, 2017). That is gradients that become extremely small or large as they are back-propagated, which may hamper the network to learn. By the same purpose, Huang et al. (2017a) proposed DenseNet. This architecture connects all the layers directly with each other to increase the flow of information throughout the network, allowing the use of fewer parameters than in previous architectures. Another CNN derived from ResNet is EfficientNet, which was proposed by Tan and Le (2019). It is characterized by a scaling coefficient that controls the network dimensions (number and size of the convolutional layers) and balances the trade-off between accuracy and efficiency.

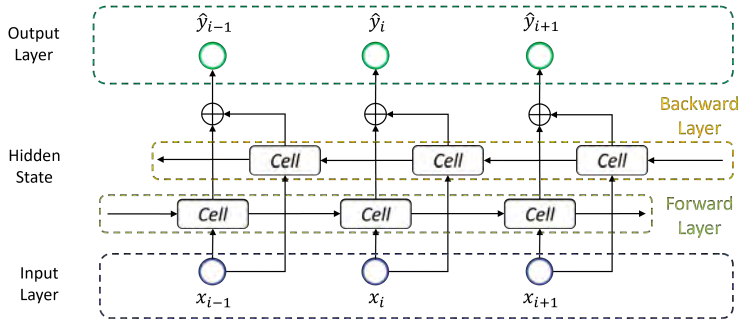
Another field where DL has shown excellent performance is in NLP, where sequences of data are processed. The first models to achieve state-of-the-art results in this field were Recurrent Neural Networks (RNNs) (Rumelhart et al., 1986). An RNN consists of a loop that allows information to be transmitted from one step to the next. This approach creates a memory of previous inputs, also called hidden states, that the network uses to produce the outputs. In other words, this type of architecture remembers previous inputs and employs that information to make future predictions. Figure 2.10a contains two schemes of a RNN. In the left part of the figure, the network is illustrated as a loop. However, it is important to note that the network is actually formed of interconnected cells or neurons, as illustrated on the right side.

RNNs are build using cells. The most basic one mainly consists of the application of an activation function over the combinations of the inputs. However, this approach has short-term memory problems due to the vanishing of the gradient. To overcome them, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) cells (Cho et al., 2014) are commonly used. Both have mechanisms to forget or remember information from the past, enabling the network to capture better long-term dependencies in sequential data in a more effective manner.

As explained so far, RNN only processes sequences in one direction. In contrast, bidirectional RNNs (Schuster and Paliwal, 1997) have two separate hidden states for each time step, so they can process the input sequence in both, forward and backward, directions. The aim of this network is to capture past and future contexts, which can be useful in certain applications, such as in language translation. To infer the outcome, the outputs from the



(a) Fold and unfold representation of a RNN structure



(b) Bidirectional RNN structure

Figure 2.10: Comparison of a RNNs and a Bidirectional RNNs.

two layers are concatenated and used to compute the final prediction. The entire process is shown in Figure 2.10b.

Recently, transformer networks (Vaswani et al., 2017) have outperformed the results obtained by RNN in NLP tasks. They are designed to achieve high performance in processing data sequences. They use self-attention mechanisms that weigh the importance of different parts of the input and make predictions based on those weights. This approach captures better the dependencies between words within a sentence compared to RNNs. Moreover, it achieves state-of-the-art results not only in language translation (Karita et al., 2019), but also in text generation (Zhang et al., 2022).

RNNs and transformers were initially designed for NLP tasks due to their ability to model sequential data effectively. However, their architectures and mechanisms can also be applied to image-related problems. RNNs can be applied to image classification by treating images as sequential data (Van Den Oord et al., 2016). On the other hand, transformers have emerged as a powerful alternative due to their ability to capture long-range dependencies efficiently (Parmar et al., 2018). These models have demonstrated promising results in image classification tasks and offer an alternative approach to traditional CNN-based architectures.

Although significant progress has been made in the field of DL, it remains a dynamic and evolving area of research with numerous opportunities and open challenges. Some of these include developing methods for explainability and interpretability, robustness, and OOD detection. If progress is made on them, DL is expected to continue playing a significant role in the development of AI and the advancement of various fields.

2.3.1 Medical Imaging and Computer-Aided Systems

Medical imaging plays an important role in modern medicine because it generates visual representations of the body using technologies such as radiography, ultrasound, endoscopy, and Magnetic Resonance Imagery (MRI) (Aiello et al., 2019). Clinicians use the data collected through these tools to identify and diagnose a wide range of lesions and diseases, monitor the progression of illnesses, guide therapeutic procedures, and evaluate the effectiveness of treatments (Chan et al., 2020a; Zhu et al., 2020; Anaya-Isaza et al., 2021; Tufail et al., 2021). These procedures are carried out manually and may be time-consuming, potentially susceptible to human error, and influenced by the subjective expertise of the reader (Brady, 2017; Waite et al., 2020). Therefore, the analysis of these medical images and the corresponding diagnosis present significant challenges, that demand extensive training and experience.

The evolution of DL has opened new horizons in the field of healthcare, making it possible to apply DNNs to a wide range of medical problems and overcome the aforementioned challenges (Wang et al., 2021). In recent years, the application of DL tools in medical imaging diagnostics has led to rapid and significant advances in CAD systems. In particular, in improving the accuracy and efficiency of the diagnostic process and helping experts to identify potential pathologies that might otherwise be missed (Chan et al., 2020b; Ostherr, 2022).

CAD systems can be divided into two types, Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx). CADe systems are designed to detect and highlight regions within an image that may indicate specific anomalies. Then, it alerts the clinician about the findings during screening. Indeed, most of the developed methods belong to this type. In contrast, CADx systems go beyond just detecting abnormalities and provide potential diagnoses, including the pre-screening and triage, evaluation of treatment response, recurrence monitoring, and prediction of prognosis or survival.

Lo et al. (1993) and Lo et al. (1995) were one of the earliest studies that attempted to use DL in the medical field. They use CNNs to detect lung nodules on chest radiography. Since then, many other studies have been conducted employing DL to detect a wide range of diseases and disorders using different images modalities, such as Computed Tomography (CT) (Higaki et al., 2019; McLeavy et al., 2021; Serte and Demirel, 2021), MRI (Lundervold and Lundervold, 2019; Mostapha and Styner, 2019; Ueda et al., 2022), and endoscopy (Hoogenboom et al., 2019; Ali et al., 2021; Wan et al., 2021). However, Chan et al. (2020a) report that although computer-assisted image analysis has become a major area of research and development, few CAD systems that incorporate advanced DL techniques have undergone extensive clinical trials, a requirement to be implemented in the medical field.

Over the past few years, there has been a strong focus on developing AI solutions that can be utilized in clinical settings (Wang et al., 2019a). Numerous promising studies

have been published, indicating the potential benefits of such technology. However, the implementation of AI in healthcare systems is currently limited due to several factors. The main one is the lack of safety studies and the absence of standardized protocols for utilizing CAD tools (Qayyum et al., 2020).

Despite these limitations, there is optimism that these challenges will be addressed through ongoing research and sufficient time. As AI systems continue improving their performance, they will become increasingly valuable in assisting healthcare professionals with tasks such as diagnosis, treatment planning, and patient care management (Yanase and Triantaphyllou, 2019b). Consequently, it is expected that DL and CAD methods will be adopted extensively in healthcare systems, ensuring their safe and reliable integration (Yanase and Triantaphyllou, 2019a).

2.3.2 Challenges in the Diagnostic

Manually diagnosing medical imaging can be lengthy and complex, as the interpretation of the images needs expert knowledge and extensive experience. To deliver a fast and high-quality reading process, it is important to leverage computer-assisted technology to ensure efficient and thorough evaluation (Allen Jr et al., 2019). However, the limitations of applying DL tools to clinical diagnostics cannot be ignored. The most common challenges when they are implemented are small and imbalanced datasets, black-box models, and OOD samples (Yanase and Triantaphyllou, 2019a; Chan et al., 2020a).

Lack of annotated data

A large amount of labeled data is needed to train DL models. However, in the medical field, the process of labeling data to train DL methods is expensive, as it is a time-consuming task and even in some cases impossible owing to the high precision required (Altaf et al., 2019). In other cases, privacy concerns associated with sharing medical records further limit the data available for training the models (Astromskė et al., 2021). Due to these facts, most studies use small training sets and over-designed systems (Vats et al., 2022). Furthermore, some of the models do not undergo strict validation using extensive test data. As a result, these models may tend to memorize the training samples instead of learning general patterns (Willemink et al., 2020). This concept is called overfitting. Hence, the generalizability of these models to new patients and clinical environments remains unclear.

Researchers often use data augmentation, transfer learning, or domain adaptation to address the lack of annotated data. Data augmentation techniques aim to artificially increase the limited data by applying transformations to images, such as rotations, flips, or adding noise (Garcea et al., 2023). In transfer learning, pretrained models are fine-tuned on smaller medical imaging datasets to leverage the knowledge of the previous method (Yu et al., 2022). Finally, the domain adaption goal is to adjust a model trained on one domain

(e.g., the source domain) to perform well in another one (e.g., the target domain) (Guan and Liu, 2022). However, these methods still face limitations and may not always yield optimal performance (Chan et al., 2020b). Therefore, it is important to find ways to improve the generalization of models with small amounts of labeled data to advance in the field of medical imaging and ensure accurate diagnoses.

Imbalanced datasets

DL algorithms work properly when datasets are balanced across different classes, i.e., all the classes have approximately the same number of samples. Unfortunately, many datasets from the medical field are imbalanced. It is a common problem because certain conditions or diseases appear in a limited number of images (Saraf et al., 2020). Thus, labeling an adequate number of abnormal samples to create an optimal dataset is a tedious task compared with obtaining normal cases, which are easier to gather for screening purposes (Akay and Hess, 2019; Yanase and Triantaphyllou, 2019a).

When the used dataset is imbalanced, the distribution of classes is skewed towards one or a few, which makes it difficult for the models to learn from minority classes. In this case, it is likely that DL systems predict the majority class (non-abnormal images) more frequently, resulting in a higher number of false negatives (missed lesions). Hence, the models over-represent the predominant class and may not learn the patterns and features associated with the minority classes. In addition, traditional evaluation metrics, such as accuracy, may not be appropriate for imbalanced datasets, because they may not reflect precisely the performance of underrepresented classes (Johnson and Khoshgoftaar, 2019).

Several techniques can be employed to address the imbalanced datasets problem, including dropout, sampling strategies, and class-balanced losses. Dropout is a technique that randomly disconnects neurons during the training of the network to improve the knowledge of others (Srivastava et al., 2014). With this, the model can prevent overfitting in the majority class and generalize more effectively. Regarding sampling strategies, different approaches exist, including oversampling the minority class (Tarawneh et al., 2022) and undersampling the majority class (Devi et al., 2020). Both strategies ensure that the model learns from sufficient examples to obtain accurate predictions. Finally, class-balanced losses can be used to adjust the importance of loss terms and mitigate the imbalance in the dataset (Cui et al., 2019; Fernando and Tsokos, 2022). In addition, the use of evaluation metrics designed for imbalanced datasets, such as the F1-score, Area Under the ROC Curve (AUC), or Mean Accuracy (MACC), can provide a more accurate and informative assessment of the model performance (Yanase and Triantaphyllou, 2019b).

Black boxes and Explainability

Complex DL models, such as CNNs, RNNs, and transformers are powerful tools for a wide range of applications. However, they present difficulties in terms of transparency and explainability as they generate outcomes without revealing how those predictions are reached. For this reason, these systems are called black boxes (Baselli et al., 2020). When DL is applied in medical applications, it is crucial to comprehend how the model arrived at its conclusions, because physicians need evidence to make or corroborate the diagnosis. Furthermore, a misdiagnosis can have a major impact on a patient's health.

Particularly, according to the study published by Zech et al. (2018), DL models may acquire knowledge of non-medical features to make predictions that are unrelated to a patient's medical conditions. These features could include aspects such as the protocols used in image acquisition, image processing techniques, or other markings and accessories that may be present in the data. These factors are often related to the facilities or coexisting conditions of the patient. Therefore, models require the ability to justify decisions to confirm that the conclusions are drawn exclusively based on medical features.

To address this challenge, approaches such as visualization techniques, attribution methods, and model interpretation frameworks, have been explored to make black boxes more reliable and transparent. Visualization techniques aim to show the internal activation and decision processes of a system (Zhou et al., 2016; Selvaraju et al., 2017; Huff et al., 2021). The goal of attribution methods is to assign importance scores to each input feature (Singh et al., 2020; Jin et al., 2022). Finally, model interpretation frameworks provide additional information that supports the understanding of the neural network outcomes (Kohoutová et al., 2020).

The field of explainable AI is in its early stages, but it is important to develop interpretable and explainable DL models to ensure their effectiveness in the tasks. In the context of medical applications, understanding and trusting the predictions made by CAD systems is crucial for their use and integration in the diagnostic processes (Linardatos et al., 2020). In addition, its comprehension helps to detect any possible biases or errors in the model, which can be corrected to enhance the performance of the system (Albahri et al., 2023).

Detection and Localization of lesions

Lesions might appear in various poses, lighting, sizes, and shapes due to the patient's anatomy and disease progression. Moreover, they are found in a small percentage of the data resulting in datasets that tend to be small and imbalanced. Such conditions make the task of detecting and localizing them consistently complex. Consequently, it is important to develop and evaluate robust CAD systems that can properly generalize and be useful for diagnosis (Mi et al., 2022).

The research community is exploring a wide range of solutions to aid in the detection

and localization of lesions. The proposed approaches go from transfer learning (Pascual et al., 2022a) or data augmentation (Xiao et al., 2022), to attention mechanisms or multi-task learning. Attention mechanisms focus on relevant regions of the image and improve the accuracy of the detection (Muruganantham and Balakrishnan, 2022), whereas multi-task learning involves leveraging shared knowledge and features across related tasks to improve performance and generalization (Vats et al., 2021).

Covariate distribution shift

The effectiveness of DL models relies on the existence of a diverse training dataset that represents accurately the intended domain or space where the system is employed. However, in the medical field, the models may have been trained using a particular version of the device and conditions, which may not align with the ones used during testing. This situation generates that data from the training and test set have different distributions, which is known as a covariate distribution shift (Nair et al., 2019; Guan and Liu, 2021).

Recent studies have proved that the performance of DL models generally deteriorates as the covariate distribution shift increase (Dharani et al., 2019). For this reason, it is essential to develop methods that are robust to changes in data distribution and that can be used with images obtained from different devices and conditions.

Several approaches have been proposed to address the disparity in data distribution shifts, including domain adaptation (Ma et al., 2019), transfer learning (Wang and Schneider, 2014), and adversarial training (Long et al., 2018). These techniques aim to bridge the gap between the source and the target domains by minimizing the distributional difference between them, thereby enabling the model to generalize better with new devices.

Out-of-Distribution data

DNNs and CAD systems are achieving incredible results in several tasks, but they still have trouble recognizing categories that they have not seen in the learning process. These examples are known as Out-of-Distribution (OOD). In any real-world application where the model may encounter unseen conditions, building robust methods to detect them is important. In the medical field is even more significant as rare images could contain abnormalities or diseases that should not remain undetected (Chen et al., 2023).

Various methods have been proposed to detect OOD data. Among them are Hendrycks and Gimpel (2016), Liang et al. (2017), and Hsu et al. (2020) which are intended to improve the model's ability to provide reliable uncertainty estimation for each image. However, OOD detection remains an open problem that requires further investigation to develop robust and generalizable solutions.

2.4 CAD Systems in WCE

The use of CAD systems based on DL techniques has the potential to revolutionize the diagnosis and treatment of digestive diseases. This new technology can ensure a more efficient and thorough evaluation of WCE studies. In particular, gastroenterologists are benefiting from the proposed and published AI techniques from the research community, which aim to assist them in the diagnosis. Some of the published work include lesion detection, video summarization, or multi-organ segmentation. All of them aim to reduce the required time to review the video, which is one of its main WCE challenges (Anaya-Isaza et al., 2021). Despite the fact that there are no standardized protocols in clinical settings for the use of CAD tools, ongoing improvement in the accuracy of these systems makes them valuable for initial screening and for more comprehensive analyses where pathologies are not immediately detected (Yanase and Triantaphyllou, 2019b).

Lesion detection is the most common approach to assist physicians during the reading process. It is a type of CAD application designed to identify any abnormality or change in the GI tract, which could indicate the presence of a lesion. These systems can identify one or multiple pathologies while reviewing each captured frame of the WCE video. New DL approaches have obtained promising results in detecting many pathologies, particularly polyp (Yuan et al., 2020; Reuss et al., 2022; Pascual et al., 2022b; Gilabert et al., 2022). In addition, the latest published systems also highlight the region where the abnormalities are located in the WCE images (Guo and Yuan, 2020; Vieira et al., 2021; Vats et al., 2021; Muruganantham and Balakrishnan, 2022).

Video summarization is a type of CAD system that condenses redundant images of one WCE study into a more manageable format. This facilitates the review and interpretation process for clinicians. Lan and Ye (2021) and Raut and Gunjan (2022) demonstrated the effectiveness of these applications by keeping the relevant medical frames from the video and reducing the time of diagnosis in WCE studies.

Other types of CAD applications in WCE that have not attracted as much attention as the previous ones are multi-organ segmentation and the proper functioning of intestinal motility. The first aims to delimit various organs by identifying the corresponding anatomical landmarks (Adewole et al., 2020; Zhao et al., 2021; Son et al., 2022). On the other hand, intestinal motility is the physiological process that coordinated the contraction and relaxation of the intestine to move food and waste through the digestive tract. Hence, the aim of these models is to determine whether this process works correctly (Malagelada et al., 2015; Seguí et al., 2016; Alcalá-Gonzalez et al., 2022).

The development of CAD systems to aid physicians in the diagnosis of digestive system disorders is an active research area, as evidenced by the numerous studies conducted thus far. By leveraging the power of DL algorithms, the accuracy, the speed of image analysis, and the detection of abnormalities in WCE videos can be potentially improved. Moreover, this advancement is expected to remove the bottleneck of human pre-reading resources and

allow for a combination of AI-validation pairs, as anticipated in recent studies (Dray et al., 2021).

The rest of this thesis is focused on continuing with the ongoing effort in the design and development of efficient and reliable CAD systems based on DL in WCE videos. Specifically, the three problems studied are pathology detection, anatomical landmark identification, and OOD sample handling.

Chapter 3

Overcoming Obstacles: Conceptual Foundations

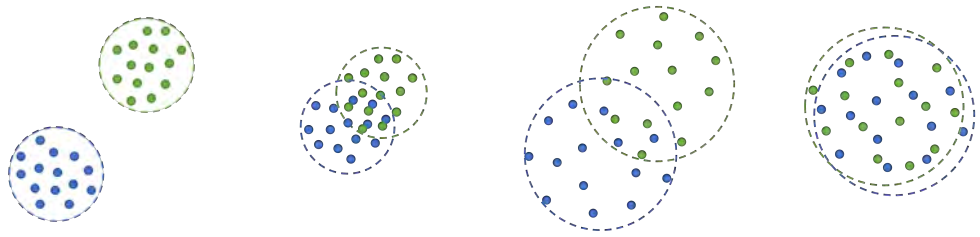
Contents

This chapter explores the theoretical fundamentals required to reach the goals proposed in Chapter 1. First, an introduction to the basic concepts and insights of DML employed in this dissertation is provided. From there, the details of the TL are explained. This loss allows the models to learn discriminative features and generalize more. Subsequently, SSL is discussed, which enables the models to learn meaningful representations from data without the need for labels. Next, the concept of the OOD problem is discussed, focusing on covariate and semantic distribution shifts, which are relevant in real-world applications, especially in the medical field. Finally, the main key takeaways of the chapter are summarized. Overall, this chapter provides a comprehensive overview of the essential concepts, literature, and techniques employed in this thesis.

3.1 Deep Metric Learning and Contrastive Approaches

Deep Metric Learning (DML) is a sub-field of DL that focuses on learning a feature space in which data that are related in some way are mapped to points close together, while data that are dissimilar are projected to points that are far apart. Consequently, DML methods can handle large amounts of data, making them suitable for real-world applications such as face recognition (Hermans et al., 2017), object detection (Xiao et al., 2020), and image retrieval (Chen et al., 2021).

The ability of DML systems to learn robust representations of the data enables the models to be insensitive to variations in inputs such as rotation, scale, and viewpoint. Hence, the models learn to detect discriminative features with low intra-class variance and high inter-class differences, as shown in Figure 3.1a. When each class forms a cluster far from the others, the networks have generalized and performed properly with test data. In contrast, if the clusters fail to exhibit low intra-class variance or high inter-class differences (Figures 3.1b - 3.1d), the absence of clear differences hampers the performance of DL models.



(a) Low intra-class and high inter-class variance.

(b) Low intra-class and low inter-class variance.

(c) High intra-class and high inter-class variance.

(d) High intra-class and low inter-class variance.

Figure 3.1: Illustration of the combinations of intra-class and inter-class variance. The points represent low representations of two different classes.

Different contrastive loss functions can be used to accomplish the objective of DML. Nonetheless, optimizing these models entails meticulous tuning of the hyperparameters and may incur significant computational costs.

Before proceeding further and explaining the most common losses used in this field, it is necessary to define some nomenclature. Let be $f_\theta(x) : X \longrightarrow Z$ a function that maps similar high-dimensional data belonging to X onto close points in the representation space Z, also known as the embedding space. Function f_θ is usually a DNN, and θ are their weights. The distance or similarity in the embedding space is computed using a metric function $D(x, y) : Z \times Z \longrightarrow \mathbb{R}$. In most cases, the Euclidean distance or cosine similarity is used. For ease of notation, $D_{f_\theta}(x, y)$ is denoted as a shortcut for $D\left(f_\theta(x), f_\theta(y)\right)$ where $x, y \in X$.

3.1.1 Contrastive Loss

Chopra et al. (2005) and Hadsell et al. (2006) were the first publications in the field of computer vision that used the contrastive loss function. The aim of this loss is to allow the network to learn how to map alike pairs of data to similar representations while ensuring that dissimilar pairs of data generate embeddings far apart from each other. In particular, the model achieved it by maximizing and minimizing the distance between pairs of samples.

Let be x_i and x_j a pair of samples from X. They are called a positive pair when they belong to the same category. In this case, the aim of contrastive loss is to reduce the distance between their representation, that is $minimize_\theta D_{f_\theta}(x_i, x_j)$. Opposite, when x_i and x_j belong to different classes, they are known as a negative pair. Now, the goal of the function is to increase the distance between them, known as dissimilarity, that is $minimize_\theta \left[m - D_{f_\theta}(x_i, x_j)\right]_+$, where $[\cdot]_+ = max(0, \cdot)$ is the hinge function, and $m > 0$ is the hyperparameter margin, which defines a radius around $f_\theta(x)$.

During the learning process, given a set of pairs of points, the contrastive loss is calculated as the sum of the distance between positive pairs and the dissimilarity of negative pairs. Besides, $y_{i,j}$ is used in the loss to specify whether the pair is positive ($y_{i,j} = 0$) or negative ($y_{i,j} = 1$). Therefore, the loss is defined as follows:

$$\mathcal{L}_{CL} = \sum_{\forall i,j} (1 - y_{i,j}) \cdot D_{f_\theta}(x_i, x_j) + y_{i,j} \cdot \left[m - D_{f_\theta}(x_i, x_j)\right]_+ \tag{3.1}$$

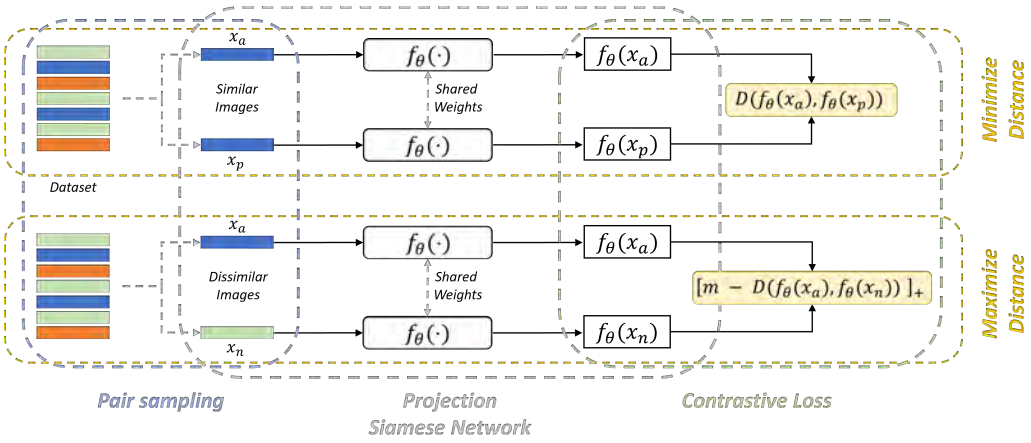


Figure 3.2: Schematic implementation of the contrastive loss using a Siamese network. The process is divided into pair sampling, data projection, and Contrastive Loss.

Training a DL model with contrastive loss involves preparing positive and negative pairs and defining a Siamese network architecture. These networks are a type of ANN structure that consists of two identical subnetworks that share the same weights. Therefore, they map each pair into the embedding space and obtain the two representations simultaneously. Then, with the obtained embeddings, the contrastive loss is computed. The scheme of this process is illustrated in Figure 3.2.

The implementation of this approach has some associated challenges. The main ones are the selection of the optimal margin, m, and mining the appropriate pairs of images, which are crucial to ensure that the learning process generates meaningful embeddings. Regardless of the method used to sample the pairs, a separate procedure from the actual training is required to calculate the distance between pairs of samples and select the most relevant. Thereby, this results in an additional computational effort that slows down the optimization process. Moreover, when large datasets are used for training, the cost of mining the right pairs quickly becomes overwhelming.

3.1.2 Triplet Loss

The Triplet Loss (TL) was first employed by Weinberger and Saul (2009) with the explicit goal of performing a k-Nearest-Neighbors (KNN) classification. Despite being an evolution of the contrastive loss, the TL takes a slightly different approach to learn rich and robust embeddings. Instead of comparing pairs of samples, the TL compares three samples simultaneously. However, it was not until 2015 that Schroff et al. (2015) proposed the current version of this loss. It uses three samples, x_a, x_p, and x_n, denoted as anchor, positive and negative. Both the anchor and positive samples belong to the same class, whereas the negative is from another category than the previous two. The aim of this loss is to ensure that the embeddings of x_a and x_p are closer than the embeddings of x_a and x_n, by at least a margin m. This constraint is formalized in Equation 3.2. In particular, Schroff et al. (2015) used Euclidean distance as the metric function $D(x, y)$.

$$D_{f_\theta}(x_a, x_p) + m < D_{f_\theta}(x_a, x_n), \quad \forall (x_a, x_p, x_n) \in \tau \tag{3.2}$$

However, to use the previous constraint as a loss, Equation 3.2 has to be reformulated as Equation 3.3. This second expression minimizes the distance between the embeddings of x_a and x_p while increasing the distance between the embeddings of x_a and x_n. The use of the margin m in this approach ensures that although all the points belong to the same class and form a single cluster, they do not collapse into a single embedding.

$$\mathcal{L}_{TL} = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} [D_{f_\theta}(x_a, x_p) - D_{f_\theta}(x_a, x_n) + m]_+ \tag{3.3}$$

Similarly to the contrastive loss, to train a neural network using the TL, it is necessary to employ Siamese networks and provide them with triplets of samples, as shown in the pipeline of Figure 3.3. Usually, the model, f_θ, quickly learns to correctly map most basic triplets, making a large portion of them ineffective. The challenge of using this loss function is that, as the dataset increases in size, the potential number of triplets grows cubically, making the training process quite lengthy. To overcome this, it is important to focus on identifying and using triplets that do not fulfill the constraint of Equation 3.2 during the

training phase. In other words, the distance between the anchor and the positive samples should be larger than the distance between the anchor and the negative samples in the embedding space.

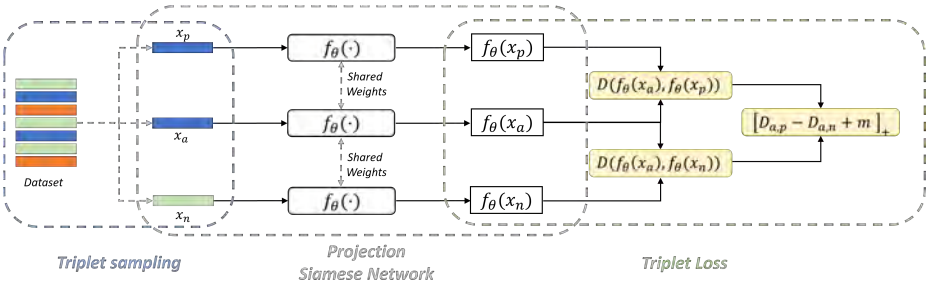


Figure 3.3: Pipeline of a Siamese network with the TL. The process is divided into triplet sampling, data projection, and the TL.

Schroff et al. (2015), Hoffer and Ailon (2015) and Hermans et al. (2017), among others, defend that mining hard triplets, i.e., $D_{f_\theta}(x_a, x_n) < D_{f_\theta}(x_a, x_p)$, is crucial for an effective learning process. Intuitively, being repeatedly informed that categories with distinct visual characteristics represent separate classes does not facilitate learning. However, observing similar-looking but dissimilar categories, (i.e., hard negatives) or images of the same class in varying poses (i.e., hard positives) can significantly aid in understanding the concept of identical categories. To find hard positives and negatives and build hard triplets, the following definitions can be used. Given an anchor image x_a, x_p is a hard positive if $argmax_{x_p} D_{f_\theta}(x_a, x_p)$. Similarly, x_n is a hard negative if $argmin_{x_n} D_{f_\theta}(x_a, x_n)$. Nevertheless, if only the most difficult triplets are presented, this results in a disproportionate selection of outliers in the data, causing the model to be unable to learn (Hermans et al., 2017). To avoid ineffective and hard triplets, it is common to mine moderate negatives (Schroff et al., 2015) and moderate positives (Shi et al., 2016), which are a middle point between easy and difficult samples.

Using the Siamese network-based approach, for each set of T triplets, T terms contribute to the loss. However, the batch contains $3T$ samples. Therefore, considering that $3T$ samples can generate up to $6T^2 - 4T$ combinations of valid triplets, only computing T is inefficient. To take advantage of all the combinations, mining strategies such as Batch All (Ding et al., 2015) and Batch Hard (Hermans et al., 2017), among others, can be used. These can improve the efficiency of the training process, by reducing the amount of memory and the computational cost required to update the model parameters. To apply these approaches, Hermans et al. (2017) propose to employ batches formed by $P \cdot K$ samples, where P is the number of random classes sampled and K is the number of samples from each class. The first strategy, Batch All, uses all possible combinations of triplets, $PK(K-1)(PK-K)$. It is formalized in Equation 3.4, where x_j^i corresponds to the j-th sample of the i-th class in the batch.

$$\mathcal{L}_{BA} = \sum_{i=1}^{P}\sum_{a=1}^{K}\sum_{\substack{p=1\\p\neq a}}^{K}\sum_{\substack{j=1\\j\neq i}}^{P}\sum_{n=1}^{K}\left[D_{f_\theta}(x_a^i, x_p^i) - D_{f_\theta}(x_a^i, x_n^j) + m\right]_+ \tag{3.4}$$

The Batch All strategy (Ding et al., 2015) requires the modification of the pipeline implementation, as it cannot be carried out with a Siamese network. To use it, only one network computes the embedding of the bs samples of the batch. Then, a pairwise distance is applied to obtain the similarity between any pair of embeddings, resulting in a tensor of size $bs \times bs$. To compute a cube tensor containing all the loss terms associated with each triplet, the second and third dimensions of the similarity matrix are expanded into two new tensors with size $bs \times bs \times 1$ and $bs \times 1 \times bs$. Then, both are subtracted to obtain the cube with shape $bs \times bs \times bs$. At that point, the margin hyperparameter, m, is added.

The resulting tensor contains the loss term for every possible combination of three images. However, certain triplets are invalid as they do not satisfy the previously mentioned constraint, which requires two images from the same class and a third from a different one. Therefore, to only consider the loss term of the valid triplets, a binary tensor mask is created using the class label information from the samples. This is done in a similar manner to the procedure employed to obtain the cube tensor. The pipeline is illustrated in detail in Figure 3.4.

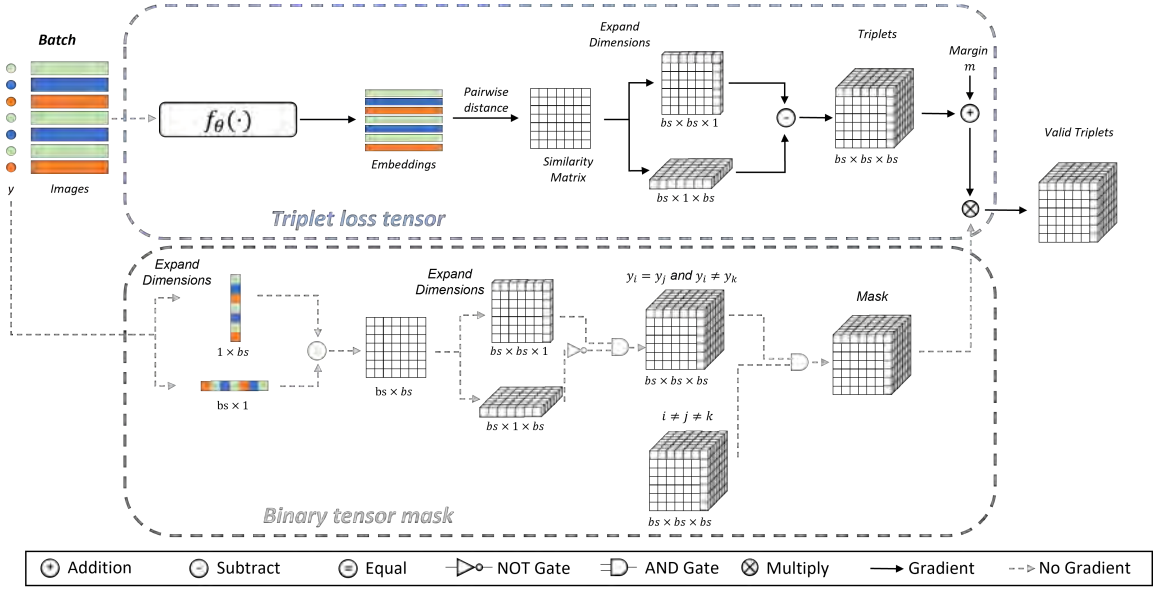


Figure 3.4: Scheme of the implementation of the Batch All strategy. The process is divided into the computation of all the triplets (top) and the creation of the binary mask to get the valid triplets (bottom).

For the Batch Hard strategy (Hermans et al., 2017), given an anchor sample x_a in the batch, the hardest positive and hardest negative samples in the batch are used to compute the loss. This strategy is formalized in Equation 3.5. Hermans et al. (2017) explained that

the obtained triplets are moderate as they are the hardest within a small subset of the data. In addition, they consider that this procedure is the best approach to optimize a network with the TL.

$$\mathcal{L}_{BH} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[m + \max_{p=1\cdots k} D_{f_\theta}(x_a^i, x_p^i) - \min_{\substack{j=1\cdots P \\ n=1\cdots K \\ j\neq i}} D_{f_\theta}(x_a^i, x_n^j) \right]_+ \tag{3.5}$$

Despite the previous online strategies being the most common ones, variants like Batch "Semi-hard", Batch "Hard and Semi-hard" or Batch "Easy, Hard, and Semi-hard" can be found in the literature. The Batch "Semi-hard" method mines semi-hard triplets, i.e., triplets that fulfill $D_{f_\theta}(x_a, x_p) < D_{f_\theta}(x_a, x_n) < D_{f_\theta}(x_a, x_p) + m$. These triplets are closer to the decision boundary than the easy ones but further than the hard ones (Schroff et al., 2015). Batch "Hard and Semi-hard" strategy selects both the hardest and semi-hardest triplets within each batch (Khaertdinov et al., 2021). Finally, Batch "Easy, Hard, and Semi-hard" approach collects easy, hard, and semi-hard triplets within each batch. It is argued that by balancing them, the model learns from all types of triplets, not only the hardest ones (Xuan et al., 2020).

Overall, TL has become a widely used loss function for training DL models. Despite its effectiveness, this loss does have a few drawbacks. The main limitations are the need for a large number of triplets and the difficulty in finding hard negatives. Multiple adaptations have been proposed to mitigate these issues, including Quadruplet loss (Chen et al., 2017b), Angular loss (Wang et al., 2017), and N-pair loss (Sohn, 2016). They extend the original TL by considering more samples, changing the comparison metric, or incorporating a different optimization objective. In several scenarios, these losses have improved the training process and the quality of the learned embeddings, thereby achieving higher performance.

It is important to note that each specific task and dataset requires the network to learn different features. Therefore, it is crucial to conduct experiments with different mining strategies and hyperparameters to determine the appropriate method and achieve the best performance.

3.2 Self-Supervised Learning

Self-Supervised Learning (SSL) is a ML technique, where the system learns from data without explicit supervision. In other words, the model is trained to perform a task without the need for ground-truth labels. Despite the concept first appearing in the early 2000s, it was not until the 2010s that it gained significant attention and made remarkable advances as it extracts information from large amounts of unlabeled data. Since then, it has been applied to several tasks, including representation learning, generative models, and reinforcement learning. The first aims to extract a compact, meaningful, and useful embedding of the input data (Ericsson et al., 2022). Alternatively, generative models create new samples that are similar to the training data (Liu et al., 2023). Finally, in reinforcement learning, the goal is to generate a policy that maximizes the reward signal (Pong et al., 2019).

Focusing solely on representation learning, the assignment performed is commonly known as a pretext task. It is specifically designed for the purpose of pretraining a model to be later fine-tuned in a downstream task. The learning pipeline is displayed in Figure 3.5. SSL methods are employed to lead to better performance and faster convergence compared with other approaches that are trained from scratch. Common pretext tasks in computer vision include image colorization (Iizuka et al., 2016), image rotation prediction (Gidaris et al., 2018), and contrastive approaches (Chen et al., 2020a). Published works concluded that SSL has the potential to reduce the amount of annotated data that is required to train DL systems, making it a valuable tool when the size of the dataset is small.

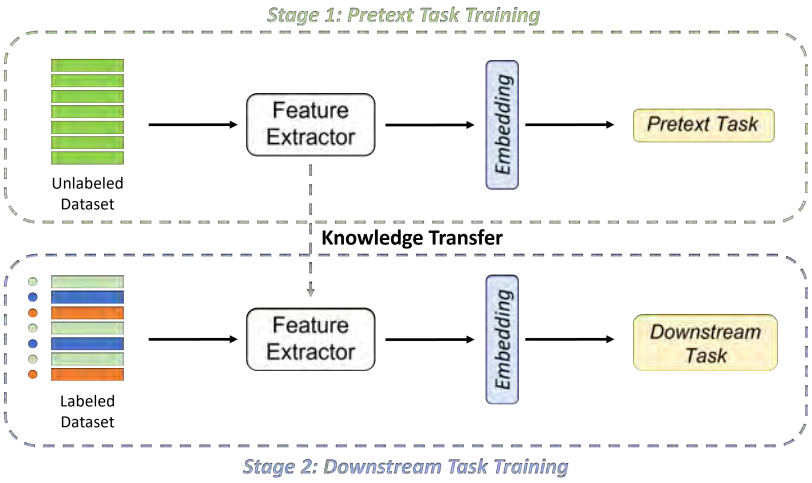


Figure 3.5: Scheme of the knowledge transfer approach. In stage 1 (top) the training is done by using a self-supervised pretext task and unlabeled data. Stage 2 (bottom) employs the knowledge learned to train the downstream task.

Rumelhart et al. (1986), Bengio (2009) and Kingma and Welling (2014) are considered the earliest works in SSL as they utilized a bottleneck architecture. Particularly, Autoencoders (AEs) (Ackley et al., 1985) are a type of ANNs that reconstructs the data

to capture the relevant features and learn compact representations. During training, AEs compresses the input into a lower-dimensional representation using a neural network as an encoder. Subsequently, another network, called a decoder, reconstructs it from this embedding. However, AEs do not obtain rich features to be applied as a pretrained model. To overcome this limitation, Bengio (2009) proposed Variational Autoencoders (VAEs), which is a variant of AEs that encodes the low-dimensional representation as a probability distribution. This modification allows the network to capture more information than with the previous approach, but sometimes it is still insufficient to perform downstream tasks.

An important breakthrough in DL and SSL was the publication of Generative Adversarial Networks (GANs), proposed by Goodfellow et al. (2014a). These systems have two networks that are trained in an adversarial manner. The first model, called generative, attempts to produce new samples that are indistinguishable from the real data. Contrarily, the second network, known as discriminative, tries to discern between the original and generated samples. The feature extraction learned by the discriminative network can be employed as pretraining for other classification tasks.

The next important milestone in SSL methods was the Colorization paper (Iizuka et al., 2016). In this study, the authors train a CNN to predict the color of grayscale images using them as the input and the ground-truth color image as the target. The network reaches its goal, producing results that are visually appealing and semantically meaningful. This procedure has been successfully implemented as a pretext task in different computer vision problems.

In 2019, Feng et al. (2019) proposed a new type of DNN architecture that incorporates rotation invariance into the feature learning pipeline. On one hand, given four rotations of an image, the network aims to predict the rotations applied to the data. On the other hand, the method penalizes the distance between embeddings obtained to ensure that they contain the same information. This methodology outperformed state-of-the-art results on standard self-supervised feature learning benchmarks.

A year later, Chen et al. (2020a) designed the Simple Framework for Contrastive Learning of Visual Representations (SimCLR). The authors proposed a three-step contrastive learning approach in which the network is trained to maximize the agreement between different views of the same image, while minimizing the agreement between the representations of different images. To that purpose, given an image x, two random data augmentations, τ_i and τ_j, are applied over x, obtaining two related views of the same image, denoted as $\tilde{x_i}$ and $\tilde{x_j}$. The applied data augmentation techniques are a random crop, followed by resizing to the original dimensions of the image, a random color distortion, and a random Gaussian blur. Subsequently, a neural network f projects both views of the image into a low-dimensional space, obtaining the embeddings $z_i = f(\tilde{x_i})$ and $z_j = f(\tilde{x_j})$. Finally, a contrastive loss function is employed to reduce the distance between the embeddings, that is, to maximize the agreement. The pipeline is illustrated in Figure 3.6.

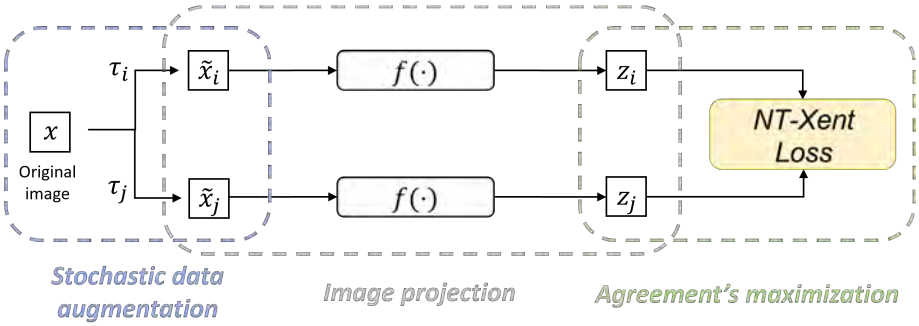


Figure 3.6: Scheme of SimCLR divided into three steps: stochastic data augmentation, image projection, and maximization of the agreement.

The contrastive loss function used by Chen et al. (2020a) is called Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) and it is expressed in Equation 3.6, where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is the indicator function, $sim(u, v) = u^T \cdot v / \|u\| \|v\|$ is the cosine similarity and T is the temperature parameter.

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/T)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(sim(z_i, z_k)/T)} \tag{3.6}$$

In the training process, the authors propose the use of N randomly selected samples in each batch. Thus, as two random data augmentations are applied, each batch contains $2N$ elements. Moreover, this method does not require explicitly sampling negative images because, for each view, the other $2(N-1)$ are used to construct negative pairs.

The results shown by the authors outperformed previous state-of-the-art SSL methods on different benchmark datasets and computer vision tasks. Consequently, SimCLR has been widely adopted by the research community as a pretext task. Modifications and extensions of the framework have been presented to improve the performance of the system. Among them, large architectures, strong data augmentations, and multi-crop evaluations have been proposed with successful results by researchers. For instance, He et al. (2020) introduced the Momentum Contrast (MoCo) framework. The key innovation was a queue of negative examples with a dynamic momentum update to train the network.

3.3 Out-of-Distribution

DL methods are built assuming that training and testing data are independent and identically distributed (i.i.d) (Liao et al., 2020). However, in real-world scenarios, this condition can hardly be satisfied, as both sets may have different ranges of feature values, and even distinct events can be found in them. When this happens, the performance of the system decreases, along with the reliability and safety of the model.

In 2017, the concept of Out-of-Distribution (OOD) samples emerged, and since then, it has gained significant research attention, resulting in a large number of methods to handle them (Yang et al., 2022). Mathematically, OOD is defined as follows. Given the joint of feature and label spaces denoted as $X \times Y$. In this space, two distributions are defined: $P_{tr}(X, Y)$ which is drawn from the training distribution, and $P_{te}(X, Y)$ which arises from the test distribution. When $P_{tr}(X, Y) \neq P_{te}(X, Y)$, implies that $P_{te}(X, Y)$ contains OOD samples.

With the given definition of OOD, three different scenarios can be defined. In the first case, the samples are categorized into the same classes but derive from different feature spaces. This particular case is called covariate distribution shift (Ben-David et al., 2010; Li et al., 2017; Wang and Deng, 2018). The second one is when the domain is the same but the categories are different, which is known as semantic distribution shift (Hendrycks and Gimpel, 2016; Liang et al., 2017; Hsu et al., 2020). And finally, the last setting occurs when the feature and the classes are distinct.

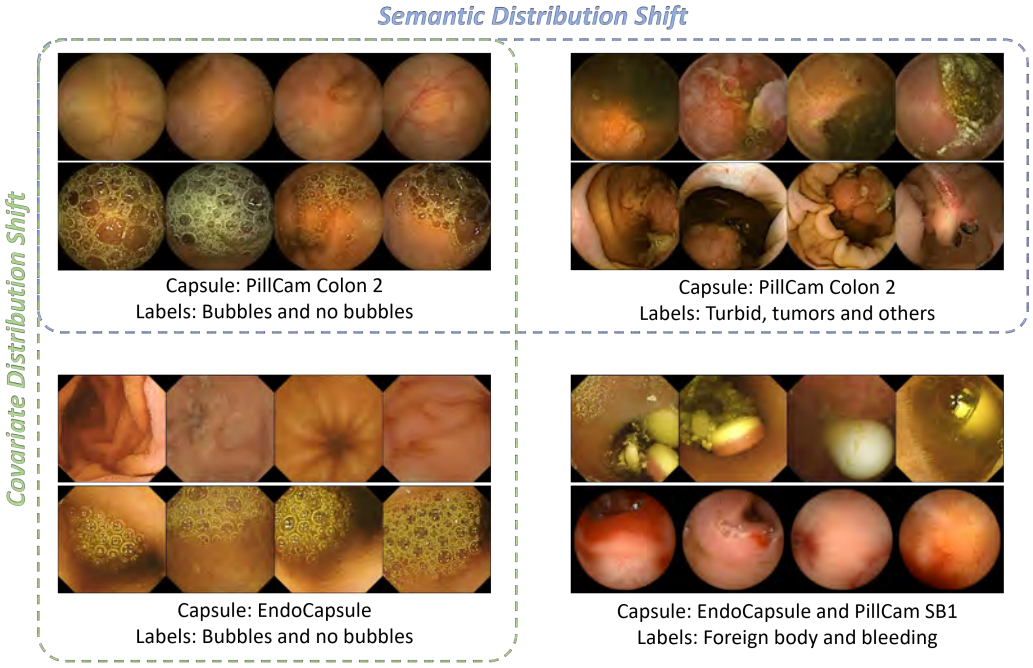


Figure 3.7: OOD classification according to the distribution shifts.

Figure 3.7 shows a visual example of the previously explained cases using WCE images. The first scenario is represented in the right column of the figure. Despite the images being labeled as bubbles and no bubbles in both cases, they are extracted from distinct capsule devices. The second case is illustrated in the first row of the figure. In it, the frames have been captured through the same capsule device, PillCam Colon 2. However, the labels are distinct. Specifically, in the second set of images, the images are labeled as a tumor, turbid, and other pathologies. The last scenario corresponds to comparing the differences in the diagonal of the figure, where the capsule and the labels are different.

Among the three scenarios that have been presented, this thesis studies the two firsts. The former is formalized as $P_{tr}(Y|X) = P_{te}(Y|X)$ but $P_{tr}(X) \neq P_{te}(X)$, meaning that the marginal distribution of X is different from training to test sets while the labels remain unchanged. In the second case, the marginal distribution of X is the same, $P_{tr}(X) = P_{te}(X)$, but the marginal distribution of Y is not, $P_{tr}(Y) \neq P_{te}(Y)$. In the following pages, more details about these two scenarios are introduced, along with their possible solutions.

3.3.1 Covariate Distribution Shift

The covariate distribution shift is addressed by aligning the feature spaces or modifying the learning process to consider the differences between the distributions. The choice of the approach depends on the specific problem, the available data, and the nature of the shift. Next, different techniques and strategies are presented.

The basic approaches for mitigating the covariate distribution shift are collecting more representative data and using preprocessing techniques. The first set of proposed methods aims to increase the quantity of data in the training phase. Several strategies can be applied, for instance, by sampling from different sources (Shrivastava et al., 2017), using synthetic images (Shafahi et al., 2019) and employing data augmentation (Garcea et al., 2023). Regarding the preprocessing techniques, common methods are batch normalization (Ioffe and Szegedy, 2015) and feature scaling. All of these approaches make the training data more diverse and reduce the impact of the distribution shift. However, these solutions only work with small variations in the data.

Transfer learning methods (Noh et al., 2019) and domain adaptation techniques (Ganin et al., 2016; Chen et al., 2018) are well-suited for mitigating the covariate distribution shift when the variation in the data is large. The former is the process of using a pretrained model to fine-tune and solve a new task that is related to the original one. A similar technique is domain adaptation, which adjusts DL systems to work effectively on data from different sources or domains. Both strategies allow the models to leverage knowledge from a domain with abundant data (source domain) to improve their performance in a domain with limited data (target domain). These approaches attempt to minimize the discrepancy between the distributions by adjusting the model parameters, re-weighting the data, or applying other techniques to generalize better to the test data. Moreover, domain adaptation can be

included in the transfer learning field.

To determine the differences between transfer learning and domain adaptation is necessary to define the concepts of domain and task. The domain refers to the feature space and its marginal distribution, whereas the task refers to the label space and the loss function. The goal of transfer learning is to transfer knowledge from one task T_a on a domain A to another task T_b in domain B, even if the domain or the task changes during the process. For domain adaptation, it is assumed that the domains and tasks remain constant. However, the marginal distributions differ between the source and target domains.

Domain adaptation can be mathematically expressed as follows. Let define $S, T \in X \times Y$, as the source and the target domain, respectively, with different distributions, $P_s(X, Y)$ and $P_t(X, Y)$. Domain adaptation aims to transfer the knowledge learned from S to T to perform a specific task on T, and this task is shared by S and T. In other words, given the source and target domain, the aim is to find an aligned space where the samples from each class, source independent, are together. This idea is illustrated in Figure 3.8.

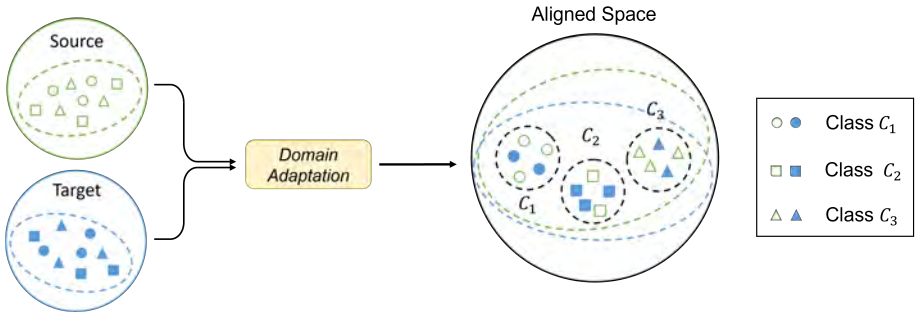


Figure 3.8: Scheme of domain adaptation.

A wide range of methods are available to perform domain adaptation. These can be classified based on their model type, label availability, modality difference, and adaptation steps (Zhao et al., 2020; Farahani et al., 2021; Guan and Liu, 2022).

Model type refers to how the features of the data are obtained. They can be categorized as shallow (Becker et al., 2015) or deep (Huang et al., 2017b). Shallow models involve human-engineered features, whereas deep ones use DL models to extract them.

Depending on the label availability, methods can be classified as supervised (Huang et al., 2017b), semi-supervised (Madani et al., 2018), or unsupervised (Ganin et al., 2016). In supervised methods, the target domain data are available to train the model. Semi-supervised systems involve a small amount of labeled and unlabeled data for training. By contrast, unsupervised models rely only on unlabeled data.

Modality differences can be divided into single-modality (Becker et al., 2015) and cross-modality (Dou et al., 2019). In the first case, data are recorded with the same type of device or source, whereas in cross-modality, data are obtained from different types of devices.

Finally, the adaptation step refers to the number of processes that the system must

perform to mitigate the distribution shift. One-step approaches are typically employed for domain adaptation (Huang et al., 2017b). However, when there is significant heterogeneity between data domains, multi-step approaches may be necessary to achieve successful alignment of the spaces (Gu et al., 2020).

3.3.2 Semantic Distribution Shift

The semantic distribution shift is addressed by building OOD detectors. These systems identify unseen events in the training set and report them to the user to avoid any consequences, particularly in real-world applications. To accomplish this, OOD detectors utilize different approaches to calculate scores or values that are used to determine whether a given image belongs to the training distribution or not. Mathematically, it can be expressed as follows. Consider a sample x belonging to the joint distribution $P_{te}(X, Y)$. Let $F(\cdot)$ represent a function that computes a score or value associated with x and let δ denote a threshold that separates the in-distribution and OOD samples. The OOD detector $g(\cdot)$ is formalized as:

$$g(x;\delta) = \begin{cases} 1 & \text{if } F(x) \leq \delta \\ 0 & \text{if } F(x) > \delta \end{cases} \tag{3.7}$$

Although there are several approaches to constructing the function $F(\cdot)$, this section specifically focuses on three types: data-only, classifier-only, and auxiliary models. However, it is important to acknowledge that OOD methods can be categorized in multiple ways, as discussed in Yang et al. (2022). Each one of them possesses its own set of strengths and weaknesses, and the choice depends on the specific task and the available datasets.

Data-only methods

Data-only approaches are based on the idea that OOD samples differ considerably from the training data. Hence, to detect them, they compare the similarity between the training and test data. To reach this goal, these methods focus on utilizing the characteristics of the data itself without requiring additional labeling or auxiliary models. However, it usually has poor generalization and needs a large amount of storage.

The simplest and most straightforward approach uses the KNN algorithm (Sun et al., 2022). To identify the OOD samples, the neighbor distance is computed between both distributions. Then, the threshold is set to determine which inputs are OOD or not.

Classifier-only methods

Classifier-only approaches measure the confidence of the model using the classification outputs. These values, called softmax scores, are obtained through the softmax function, which is the last activation of the output layer, where logits are converted into probabilities. Classifier-only methods operate on the principle that well-trained models assign higher softmax scores to in-distribution data than to OOD samples. These approaches do not require any modification in the training procedure. Moreover, contrary to data-only methods, classifier-only models require less stored data.

One of the first classifier-only methods was Maximum over Softmax Probabilities (MSP) (Hendrycks and Gimpel, 2016). It utilizes the softmax score from the classifier trained on $P_{tr}(Y|X)$ to detect OOD samples. Despite the low efficacy of the method due to its simplicity, it acted as a stepping stone for subsequent developments. One year later, Liang et al. (2017) proposed Out-of-Distribution Detector for Neural Networks (ODIN), an improvement of MSP. This method is a confidence-based approach based on two components: temperature scaling (Geoffrey Hinton and Dean, 2015) and input preprocessing (Goodfellow et al., 2014b). Using them, the authors aim to achieve highly peaked softmax outputs for in-distribution samples and flat ones for OOD samples. Further details of these two components are given next.

Generally, the softmax activation function produces very sharp probability distributions, which implies that the model is overconfident in its predictions, even when the input is slightly different from those from the training. To reduce the confidence in the model's outcomes and make it more sensitive to small changes in the input, Geoffrey Hinton and Dean (2015) proposed a technique called temperature scaling. When it is applied to the logits before the softmax function, the entropy or degree of uncertainty of the probability distribution is increased. To formalize this method, let be x the input and $f_\theta = (f_1, \cdots, f_n)$ a neural network trained to classify n classes. The output predicted by the neural network is $\widehat{y}(x) = argmax_i S_i(x;T)$, where $S_i(x;T)$ is the calibrated softmax score and is defined in Equation 3.8, being $T \in \mathbb{R}^+$ the temperature scaling parameter.

$$S_i(x;T) = \frac{\exp(f_i(x)/T)}{\sum_{j=1}^{N} \exp(f_i(x)/T)} \tag{3.8}$$

It is noteworthy that during the training phase, T is set to 1. Finally, the softmax score is the maximum softmax probability, that is, $S_{\widehat{y}}(x;T) = \max_i S_i(x;T)$.

Alternatively, the input preprocessing in ODIN is inspired by Goodfellow et al. (2014b). They discussed that preprocessing an image by introducing small perturbations decreases the softmax score of any given input. Moreover, it exposes potential weaknesses in the network and makes it easy to detect OOD samples. Essentially, to create these perturbations, it is enough to add small amounts of noise or distortions in all the data. In the case of ODIN, the input preprocessing is formulated in an adversarial manner as displayed in

Equation 3.9, where ϵ is the perturbation magnitude parameter.

$$\tilde{x} = x - \epsilon \cdot sign(-\nabla_x \log S_{\widehat{y}}(x; T)) \tag{3.9}$$

To determine if an input is OOD sample, ODIN combines the two previously detailed components, the scaling temperature and the input preprocessing. First, preprocessing is applied over x to obtain $\tilde{x}$. Subsequently, the calibrated softmax score $S(\tilde{x}; T)$ is computed using f_θ. Finally, the obtained score is compared against the threshold δ to determine if the evaluated image is an OOD sample. Mathematically, the function that formalizes the detection of OOD samples is:

$$g(x; \delta, T, \epsilon) = \begin{cases} 1 & \text{if } max_i S_i(\tilde{x}; T) \leq \delta \\ 0 & \text{if } max_i S_i(\tilde{x}; T) > \delta \end{cases} \tag{3.10}$$

As reported by Liang et al. (2017), the hyperparameters T, ϵ, and δ have to be selected to correctly classify the 95% of the in-distribution images.

Finally, another classifier-only method is Hsu et al. (2020). It is based on ODIN but without requiring any OOD data for tuning. For it, the authors propose two strategies to improve the detection performance. The first one involves decomposing the confidence scoring process, while the second strategy modifies the hyperparameter ϵ of the input preprocessing method. These modifications allowed the model to achieve state-of-the-art results, outperforming ODIN.

Auxiliary models

Auxiliary approaches attempt to learn what a training distribution sample looks like. Particularly, to identify OOD samples, they evaluate the performance of the model on auxiliary tasks. The intuition is that if the model performs poorly on them, then it is likely that the input is OOD.

These methods can be implemented using deep-generative models such as AEs (Ackley et al., 1985) and VAEs (Bengio, 2009; Kingma and Welling, 2014). Intuitively, these models are optimized to reconstruct the training data. However, when OOD data are reconstructed, their quality is expected to be poor and with a higher error, which allows defining a threshold to detect them.

3.4 Key Takeaways

Although DL might be a powerful tool for multiple applications, it also has a significant number of challenges, including imbalanced datasets, lack of annotated data, and the existence of OOD samples. The previous sections of this chapter have detailed the advanced technologies that will be used in this thesis to overcome WCE diagnosis and DL challenges. Next, the key takeaways are summarized below.

The TL is a DML technique for learning better representations, particularly in the context of image recognition tasks. This loss facilitates the learning of low-dimensional representations of images, ensuring that related images are embedded close to each other, while dissimilar images are projected far apart. This can encourage the model to discriminate better between classes and capture more detailed information in the learned representations. In addition, it can mitigate the effects of imbalance and small datasets. However, generating all the possible triplets to train the network is not completely effective. Therefore, batch sampling strategies that involve mining triplets of anchor, positive, and negative samples within a batch might be one plausible strategy to overcome the TL limitations.

SSL is an emerging paradigm that has shown great promise in leveraging unlabeled data. The basic idea is to define pretext tasks that encourage the model to learn useful representations without explicit supervision. SSL can yield highly generalizable and transferable features that can be applied to a wide range of downstream tasks. The success of these methods has been further enhanced by recent developments in contrastive learning approaches. As a result, its application to obtain pretrained models might improve the performance in downstream tasks, even with the presence of imbalanced and small datasets.

Domain adaptation is a powerful technique for addressing the covariate distribution shift. The main idea is to leverage the similarities between the source and target domains to effectively transfer knowledge from the first to the second one. This can improve the model's ability to generalize to new data.

One of the major challenges in DL is the impact of OOD data, which can lead to catastrophic errors in model predictions, especially in the healthcare field. Therefore, OOD detection is a critical technique to prevent further complications. Particularly, ODIN has shown hopeful results in addressing this issue. The key idea is to enhance the model's uncertainty estimation by using temperature scaling and input preprocessing. By doing so, ODIN can distinguish between in-distribution and OOD samples, leading to improved model robustness and reliability.

As has been previously shown, leveraging these advanced techniques enhances the robustness, reliability, and generalizability of the models. It is important to note that the choice of technique depends on the specific problem and application at hand. However, by incorporating these techniques into our DL pipelines, the performance and effectiveness of diagnostics can be improved.

Chapter 4

Paper I

WCE polyp detection with triplet based embeddings

Pablo Laiz[a], Jordi Vitrià[a], Hagen Wenzek[b], Carolina Malagelada[c], Fernando Azpiroz[c] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

c) Digestive System Research Unit, University Hospital Vall d'Hebron, Barcelona, Spain

In Computerized Medical Imaging and Graphics, Elsevier, Volume 86:101794, 2020. ISSN 18790771.

doi: 10.1016/j.compmedimag.2020.101794

Impact Factor Journal: 4.790
Q1 Biomedical Engineering (22/98)

Contents

Motivation and Context

Colorectal polyps are a common and potentially dangerous condition found in the GI tract. To prevent further health complications such as CRC, it is crucial to detect them in the early stages. A screening technique used to identify polyps is WCE. However, identifying them is challenging, even for experienced gastroenterologists. Polyps appear in only a few frames in up to 8 hours of videos and can vary in position, morphology, and size. Therefore, analyzing frame-by-frame a WCE video is a time-consuming and error-prone task (Vasilakakis et al., 2019; Yang, 2020). To address it, DL-based CAD systems are well-situated to assist physicians (El Ansari and Charfi, 2017; Rahim et al., 2020).

In the following study, instead of using a traditional classification approach, a novel CAD system is proposed and evaluated as an information retrieval problem. The model aims to rank the images of a WCE video and display the most relevant at the beginning of the list of results shown to the physician. This strategy facilitates the screening of a smaller set of images and maximizes the capability of detecting polyps.

The development of CAD systems for polyp detection is difficult. First, the datasets used to train the models have an imbalanced structure Akay and Hess (2019). It means that the number of positive images (polyps) is significantly smaller compared to negative samples. The second problem is the scarcity of data, as obtaining a large number of annotated WCE images is a lengthy and costly process (Vats et al., 2022). To overcome the challenges of

small and imbalanced datasets, the proposed CAD system employs the TL to optimize the model. This loss function pushes the embedding of images from the same class together and maximizes the distance between the images from different classes, resulting in improving the feature extraction process and obtaining robust representations.

In the current literature, the evaluation metrics used for polyp detection in WCE images do not always reflect the actual performance of CAD systems (Johnson and Khoshgoftaar, 2019). In this paper, a set of recommendations and guidelines are proposed to show their real performance, taking into consideration the imbalanced dataset structure and the scarcity of data.

The experiments conducted in this study demonstrate that the proposed CAD system outperforms both the baseline and state-of-the-art methods for polyp detection in WCE images. The superior performance of the model is attributed to the use of the TL. Overall, this study highlights the importance of developing CAD systems for polyp detection and presents a new approach to address the challenges associated with this task.

Furthermore, for the effective implementation of this CAD system in clinical practice, it is important to increase the trust and acceptance of physicians in it. For this reason, this paper includes a section with qualitative results, where the method proposed by Zhou et al. (2016) is used to explain the reason behind the model's predictions. It enables clinicians to understand in which part of the image the polyp is located, thus increasing their confidence in the system.

Finally, it is important to note that the method presented in this paper has been used in the clinical trial Lei et al. (2023b), to evaluate its effectiveness in detecting polyps as an AI-reader pair. This trial aimed to gather empirical evidence to support the use of this system in medical practice.

WCE polyp detection with triplet based embeddings

Pablo Laiz[a], Jordi Vitrià[a], Hagen Wenzek[b], Carolina Malagelada[c], Fernando Azpiroz[c] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

c) Digestive System Research Unit, University Hospital Vall d'Hebron, Barcelona, Spain

4.1 Abstract

Wireless capsule endoscopy is a medical procedure used to visualize the entire gastrointestinal tract and to diagnose intestinal conditions, such as polyps or bleeding. Current analyses are performed by manually inspecting nearly each one of the frames of the video, a tedious and error-prone task. Automatic image analysis methods can be used to reduce the time needed for physicians to evaluate a capsule endoscopy video. However, these methods are still in the research phase.

In this paper, we focus on computer-aided polyp detection in capsule endoscopy images. This is a challenging problem because of the diversity of polyp appearance, the imbalanced dataset structure, and the scarcity of data. We have developed a new polyp computer-aided decision system that combines a deep convolutional neural network and metric learning. The key point of the method is the use of the Triplet Loss function with the aim of improving feature extraction from the images when having a small dataset. The Triplet Loss function allows training robust detectors by forcing images from the same category to be represented by similar embedding vectors while ensuring that images from different categories are represented by dissimilar vectors. Empirical results show a meaningful increase in AUC values compared to state-of-the-art methods.

A good performance is not the only requirement when considering the adoption of this technology to clinical practice. Trust and explainability of decisions are as important as performance. With this purpose, we also provide a method to generate visual explanations of the outcome of our polyp detector. These explanations can be used to build a physician's trust in the system and also to convey information about the inner working of the method to the designer for debugging purposes.

The paper has been re-typeset to match the thesis style.

Key words: Deep metric learning; Triplet loss; Deep learning; Capsule endoscopy; Polyp detection.

4.2 Introduction

According to the Global Health Organization, colorectal cancer is the third most frequent type of cancer with 1.8 million people diagnosed in 2018 (Siegel et al., 2019). The early detection of cancer, when it is still small and has not spread, is essential for the treatment and the survival of the patient. The detection and removal of intestinal polyps, an abnormal growth of the tissue that can evolve into cancer, is especially important. According to the American Cancer Society, screening tests of the Gastrointestinal (GI) tract have significantly increased the survival rate of colorectal cancer patients*.

The standard clinical procedure for screening the rectum and the colon is a colonoscopy. Despite the fact that this procedure is widely accepted, it has some important drawbacks: it requires qualified personnel, expensive medical facilities and may result in patient discomfort.

Wireless Capsule Endoscopy (WCE), originally developed by Iddan et al. (2000), is an alternative technique designed to visualize the inside of the digestive tract with minimal patient discomfort. Patients ingest a vitamin-size capsule that contains a camera and an array of LEDs powered by a battery, to record and send the captured images to an external device for a posterior analysis.

WCE has become a solution to the rapid increase in demand for optical endoscopies in recent years (Li and Leung, 2018), as it can deliver GI investigations without the need for expensive clinical resources and much improved patient comfort. It has been reported report that this device can accurately evaluate pathologies such as gastrointestinal bleeding (Usman et al., 2016; Jia and Meng, 2016; Zwinger et al., 2019), Crohn's disease (Goran et al., 2018), ulcerative colitis (Ozawa et al., 2019; Maeda et al., 2019), small-bowel tumors, polyposis syndrome (Byrne and Donnellan, 2019; Yang, 2020) and is also applicable in polyp detection (Kobaek-Larsen et al., 2018). Furthermore, recent studies (McGoran et al., 2019; Takada et al., 2020) foresee WCE as a tool to not only investigate the symptoms of GI disorders but also, in the future, to perform therapeutic interventions. The capsule could also democratize the screening process since it is better tolerated than standard endoscopy (McGoran et al., 2019), has minimal invasiveness with user-friendliness (Vasilakakis et al., 2019), does not require sedation and has fewer potential complications.

Although WCE presents many advantages over other screening techniques, it presents an important drawback in clinical practice: resulting videos can contain hundreds of thousands of images per patient that must be screened by clinical specialists. This screening is complex, tedious and time-consuming, often lasting 2 to 3 hours per video (Vasilakakis et al., 2019;

*https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html

Yang, 2020). Moreover, and also because of the fatigue caused by the visual inspection of these videos, it is common to review procedures more than once to ensure that no pathological images are missed (Byrne and Donnellan, 2019). All these inconveniences hinder the adoption of WCE, exposing the need of Computer-Aided Detection and Diagnosis (CAD) support systems (El Ansari and Charfi, 2017; Rahim et al., 2020) with artificial intelligence (McGoran et al., 2019; Hwang et al., 2018; Yang, 2020).

In the literature, we can find several AI-based CAD systems specially designed to detect suspicious or abnormal WCE images. Most of these methods are aimed at reducing visualization and diagnosis time by detecting specific GI events with high-performance machine learning systems (Takada et al., 2020).

With regard to the specific goal of polyp detection, most of the published systems have been reported and validated as automatic detection methods. However, because of legal and practical reasons, these systems cannot be used for automatic diagnosis and can only be deployed as decision support systems that filter the whole set of frames to allocate physician's attention to those images that show potential polyp structures. In most cases, this is a needle-in-haystack problem because of the occasional appearance of images with these pathologies. Figure 4.1 shows two sequences from different procedures where a polyp is observed. It is important to point out that, in both procedures, those are the only images of the whole procedure where a polyp is visible. Figure 4.2 shows some random images from the same procedures.

Figure 4.1: Illustration of two polyp sequences extracted from different patients. In the first sequence can be seen how the polyp appears in all the frames approximately in the same location. However, in the second sequence, the polyp location changes while the WCE moves through the GI tract.

Polyp detection has been an active research topic, as can be seen in Table 4.2. However, to our knowledge, there is no agreement about a common evaluation methodology to allow the community to compare different CAD methods. Most of these methods have been developed and evaluated with private datasets and using different evaluation methodologies, which are suited for image detection systems but not fully informative for CAD systems in medical applications.

In this paper, we propose and validate a CAD system for the detection of polyps in WCE

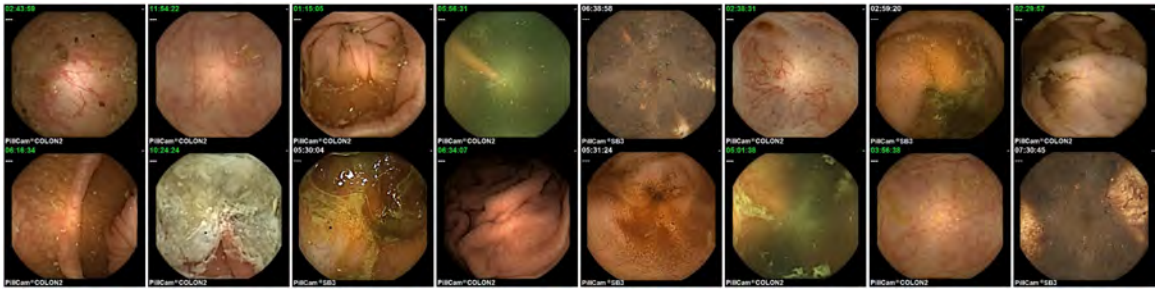


Figure 4.2: Illustration of 16 random samples obtained from the same procedures that represent the huge diversity of the dataset. For example, some of the frames present turbid, GI walls or wrinkles among others.

videos. The proposed system is based on deep Convolutional Neural Network (CNN). It is well known that CNNs have become state-of-art in many visual recognition problems, but their application in the medical field has been rather limited, with some exceptions like dermatology and breast x-rays. The main reason for this is that medical databases are comparably poor and small due to the high costs involved in data acquisition, their complex labeling, and because the use of these data often involves confidentiality issues (Akay and Hess, 2019).

Small size and imbalanced data are two of the main obstacles to develop reliable Deep Learning (DL) classifiers, because if not properly addressed, they may lead to training overfitting and poor generalization. Several tricks and techniques, such as dropout (Srivastava et al., 2014), sampling strategies (Katharopoulos and Fleuret, 2018), image augmentation (Mikołajczyk and Grochowski, 2018) and curriculum learning (Taghanaki et al., 2019), try to alleviate this problem, but it is still an open and important challenge in the medical field as described in Akay and Hess (2019). To this aim, and to overcome the small amount of available data for training the CNN, in this paper we propose an optimization strategy based on deep metric learning that uses the Triplet Loss function (Schroff et al., 2015). The obtained results show that this learning strategy outperforms the classical learning strategy using the cross-entropy loss function in our problem.

Our contributions are as follows:

- We propose an evaluation methodology that involves quantitative metrics as well as the reporting of qualitative database information in order to allow fair comparisons between different systems.
- We show how to build an end-to-end CNN polyp detection system, based on the Triplet Loss function, that overcomes the problem of imbalanced datasets.
- Finally, we propose the use of classifier interpretation techniques as a mechanism to build trust in the system.

The paper is organized as follows: first, we give an overview of the related work in the

field. This is followed by a description of our methodology, presenting the system architecture, parameter optimization and evaluation methodology, followed by experimentally setup and results. Finally, we conclude the paper and give directions for future work.

4.3 Related Work

Since the presentation of WCE, several computational systems have been proposed to reduce its inherent drawbacks in clinical settings (Liedlgruber and Uhl, 2011; Belle et al., 2013). Generally, these systems are designed either for efficient video visualization (Mackiewicz et al., 2008; Chu et al., 2010; Iakovidis et al., 2013; Drozdzal et al., 2013) or to automatically detect different intestinal abnormalities such as bleeding (Jia and Meng, 2016; Usman et al., 2016; Li et al., 2017), polyp (Yuan et al., 2018; Zhang et al., 2019), tumor (M Cobrin et al., 2006), ulcer (Ciaccio et al., 2013), motility disorders (Malagelada et al., 2015; Segui et al., 2014) and other general pathologies (Ciaccio et al., 2010; Kumar et al., 2012; Malagelada et al., 2012; Chen et al., 2013; Zhao et al., 2015). DL nowadays represents the state-of-the-art to most of these problems. Table 4.1 shows detailed information on those systems that have been implemented using DL methods.

Table 4.1: Comparison of existing DL methods for the classification problem in WCE. In the last column, Metrics, the legend used is: Accuracy (A), Sensitivity - Recall - TPR (B), Specificity - TNR (C), ROC (D), AUC (E), Precision (F), Confusion Matrix (G), F1-Score (H), Cohen's Kappa score (I).

Reference (Year)	**Class**	**Dataset**		**Validation**		**Architecture**	**Metrics**
		Videos	Images	Method	Patient Separation		
Zou et al. (2015)	Localization	25	75k	60k / 15k	Unknown	AlexNet	A
Yu et al. (2015)	Digestive organs	25	1M	60k / 15k	Unknown	CNN + ELM	A
Seguí et al. (2016)	Scene classification	50	120k	100k / 20k	Unknown	CNN	A-G
Jia and Meng (2016)	Bleeding	-	10k	8.2k / 1.8k	Unknown	AlexNet	B-F-H
Li et al. (2017)	Haemorrhage	-	11.9k	9.6k / 2.24k	Unknown	LeNet, AlexNet, GoogleNet, VGG-Net	F-B-C-H

Among possible WCE uses, polyp detection has been one of the problems that have attracted a lot of attention from researchers. Table 4.2 presents a set of methods, published in high-impact conferences and journals, aimed at detecting polyps by using any of the GI examination modalities. As can be seen, prior to 2015, most of the published methods were based on conventional computer vision and machine learning techniques, which are based on the extraction of handcrafted visual features followed by a classifier. These systems have used several image features such as color, texture, and shape to deal with the classification task.

Since 2015, and following the success of DL methods in any computer vision application, most of the proposed methods to detect WCE events are based on DL. One of the first methods, known as SSAEIM, was proposed by Yuan and Meng (2017). This method,

Table 4.2: Overview of our proposed and existing method for polyp detection. The nomenclature is the same as in Table 4.1.

Reference (Year)	Modality	Dataset			Validation		Feature	Metrics
		Videos	Polyp	Non-polyp	Method	Patient Separation		
Li et al. (2009)	WCE	2	150	150	3-fold	Unknown	Colour and shape	A-B-C
Zhao and Meng (2011)	WCE	2	-	-	5-fold	Unknown	Colour	A-D-E
Li and Q.-H. Meng (2012)	WCE	10	600	600	10-fold	Unknown	Texture	A
Yuan and Meng (2014)	WCE	10	436	436	10 random splits	Unknown	Texture	A-B-C
Bae and Yoon (2015)	Endoscopy	141	1k	100k	5-fold	Unknown	Shape	B-E-F
El Khatib et al. (2015)	Colonoscopy	20	2k	3k	-	Unknown	Texture	B
Zhu et al. (2015)	Endoscopy	-	6.5k	50k	10-fold	Unknown	CNN	A-B-C
Zhang et al. (2017)	Colonoscopy	-	826	1.1k	Random test	Unknown	CNN	A-C-F-H
Yuan et al. (2017b)	Colonoscopy	6	37k	36k	Random test	Unknown	CNN	A-B-G
Yu et al. (2017)	Colonoscopy	20	3.7k	-	18 dif. videos	Separate	FCN	G-F-B-H
Yuan and Meng (2017)	WCE	35	-	-	1k /3k	Unknown	SAE	A-G
Yuan et al. (2018)	WCE	62	1.5k	1.5k	150/150	Unknown	CNN	A-B-F-H
Zhang et al. (2019)	Colonoscopy	215	404	-	50 rand. images	Unknown	CNN	G-F-B-H
Yuan et al. (2020)	WCE	80	1.2k	6k	120/600	Unknown	CNN	A-B-C-F-H
Guo and Yuan (2019)	WCE	-	585	2.2k	4-fold	Unknown	CNN	A-I
This Paper	WCE	120	2.1k	1.3M	5-fold	Separate	CNN	

which uses a set of concatenated sparse autoencoders and a reconstruction loss to automatically find powerful features for the classification task, shows an improvement over previous methods based on handcrafted features.

Yuan et al. (2018) argued that object rotation and high intra-class variability are two main limitations for WCE image analysis. In order to overcome this problem, the authors proposed a new method named RIIS-DenseNet, based on a DenseNet, which uses two loss functions as constraints. The rotation-invariant constraint was designed to achieve rotation invariance by enforcing similarity between feature representations of the training samples and their corresponding rotated ones. Meanwhile, the image similarity constraint was proposed to allow a small intra-class variance in the feature space.

The same authors, Yuan et al. (2020), proposed DenseNet-UDCS one year later, aiming to overcome three different problems: an imbalanced database, small inter-class variance, and high intra-class variability. To achieve these goals, the network uses a weighted cross-entropy loss together with a category sensitivity loss. Weighted cross-entropy is appropriate to deal with the imbalanced dataset while category-sensitive loss aims to reduce the distances between feature representation of samples from the same category.

Guo and Yuan (2019) proposed Triple ANet, a CNN system which addresses the problems of high intra-class variability, small inter-class variance, and the existence of artifacts in the images. The system introduces two blocks of deformable convolutions to capture correlations and highlight informative areas in the images. The other fundamental point of the system is the replacement of the classifier by an angular contractive loss.

These methods represent a significative achievement, overcoming some of the main problems related to WCE image analysis, but we think that the solution to this problem is not complete if the right metrics and evaluation methodologies are not used for validating them. There are three features of these methods that are worth analyzing in order to define the

right comparison methodology: database size, validation strategy, and evaluation metrics.

Databases: As it can be seen, in most of the cases the number of polyps in the dataset is relatively small. The paper that uses the largest number of polyps uses a total of 37,000 images, while the smallest uses just 25. If we consider only those papers which work exclusively with WCE images versus also colonoscopy images, the number of images is significantly smaller. The paper that uses the largest dataset uses a total of 1,500 polyp images obtained from 62 different patients, which means an average of 25 polyp frames per procedure. It is also important to point out that the number of procedures is 2 to 1,000 times smaller than the number of polyps. This means that several images from the same patient or even polyp are used in the dataset, but this information is not usually reported. Besides this and in order to understand how challenging the dataset is, it would also be important to report the size and type of polyps. Regarding negative samples, the paper that uses the largest databases uses a total of 100,000 images while the paper that uses the smallest set uses 75 images. No information about the sampling strategy that was used to obtain these negative images is reported in any case.

Training and validation strategy: As pointed out before, datasets usually contain several positive images from the same patient, and in most cases several images from the same polyp. For this reason, it is very important to ensure that the training and the validation sets do not share images from the same procedure. If the partition of the training and validation is not properly done, it would be highly probable to have consecutive and practically identical frames in both sets. This fact clearly contaminates any validation result based on those datasets. Only the method presented by Yu et al. (2017) reports this information.

Evaluation metrics: In order to validate these systems, authors use a variety of evaluation metrics: accuracy, precision, sensitivity/recall, specificity, ROC-Curve, AUC, F1-Score as well as the confusion matrix. The diversity of evaluation metrics clearly hinders a clear comparison between methods, thus showing the need for a good and unified evaluation strategy.

Taking into account that we are dealing with a computer-aided decision system, we designed and evaluated our approach not as a classical classification problem, but as an information retrieval problem. Given a WCE video, the problem is to rank the images of the video according to some criterion so that the more relevant images appear early in the result list displayed to the physician. This allows the visual screening of a reduced set of images and at the same time ensures the detection of a maximum number of positive images.

The description of the method is divided into the following three parts:

- System Architecture: Introduction of the CNN architecture used to detect polyp images.

- Parameter Optimization: Explanation of how the chosen architecture is optimized. Presentation of the problems derived from the database and how to adapt the learning process to achieve better results.
- Evaluation Methodology: Presentation of standard metrics and discussion about how to evaluate polyp detection systems to be able to compare them with other works.

4.3.1 System Architecture

The proposed DL method is based on the Deep Residual Network (ResNet) architecture, presented by He et al. (2016). This network has shown outstanding results in important image recognition problems.

The main novelty of this architecture is the use of a high number of layers that progressively allow to learn more complex features. The first layers learn edges, shapes or colors while the last ones are able to learn concepts. In order to learn, this architecture needs the introduction of a set of *residual blocks* that avoid the problem of vanishing gradients when the number of layers increases. These blocks are built by using skip connections or identity shortcuts, that reuse the outputs from previous layers.

The residual block has the following form:

$$y = F(x, \{W_i\}) + x \tag{4.1}$$

where $F(x, \{W_i\})$ represents stacked non-linear layers and x the identity function.

Taking into account the performance of this architecture in other image classification problems, we used the fifty-layer ResNet version, known as ResNet-50.

4.3.2 Parameter optimization

ResNet-50 has over 23 million trainable parameters. The robust estimation of these parameters needs millions of images as described in He et al. (2016), but these parameters have been shown useful for a variety of visual recognition problems. The original paper used the cross-entropy loss function with an L2 regularization term to estimate all these parameters. Binary cross-entropy loss function decreases as the prediction converges to the true label, and increases otherwise, as its function indicates:

$$L_{CE}(p, y) = -y \cdot \log(p) - (1 - y) \cdot \log(1 - p) \tag{4.2}$$

where y is the true label of the sample and p is the estimated probability of the sample belonging to class 1.

In our case, to deal with a small and imbalanced dataset, we propose an optimization of the ResNet in two stages. First, images are projected into an embedding space using the Triplet Loss (TL) as described in Schroff et al. (2015). Then, we consider the cross-entropy

loss function in the embedding space. The proposed methodology is shown in the upper part of Figure 4.3.

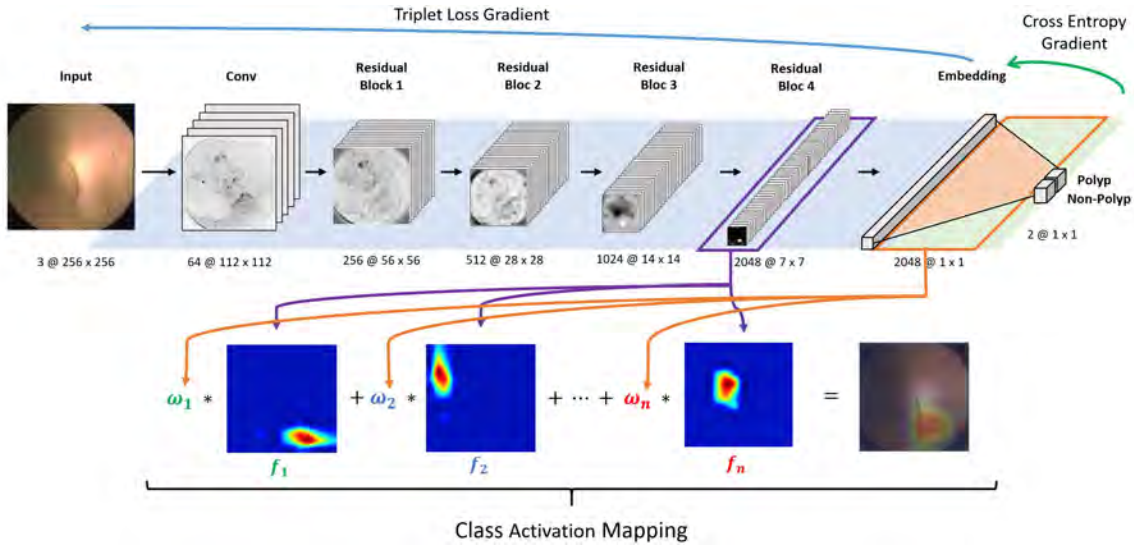


Figure 4.3: Overview of the proposed CNN structure. The upper part of the scheme appears the ResNet architecture with our methodology applied to it. The background color reflects the layers affected by each one of the gradients generated by the main losses. The lower part of the figure shows how the class activation mapping is built.

TL, a Deep Metric Learning (DML) method, has shown great generalization results when dealing with a large number of classes as for instance in the problem of face identification. The goal of TL is to optimize a latent feature space $f(x) \in \mathbb{R}^d$ such that examples from the same class are closer to each other than to those belonging to different classes.

In order to learn this embedding representation, TL aims at ensuring that an anchor image, x_a, is closer to all other images from the same class, x_p, than any image from a different class, x_n. This concept, illustrated in Figure 4.4, can be formulated as follows:

$$\|f^a - f^p\|_2^2 + \alpha < \|f^a - f^n\|_2^2, \forall (x_a, x_p, x_n) \in \tau \tag{4.3}$$

where f^k is the embedding of $f(x_k)$, $\|\cdot\|_2$ is the Euclidean distance and α is a margin, which defines the minimum distance between elements of different classes.

In order to train the network and reach the sought condition, the TL function is defined as follows:

$$L_{triplet-loss} = \sum_{i=1}^{N} \Big[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha \Big]_+ \tag{4.4}$$

Training a neural network using TL is not simple. At training time, the network receives triplets of images. For small datasets, the generation of each triplet is feasible, but when the amount of images increases, the number of possible triplets grows with cubic complexity. If we try to generate all of them, it becomes intractable and inefficient, making it impossible

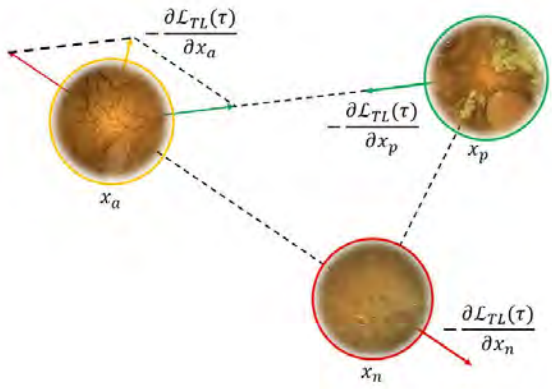


Figure 4.4: Behavior representation of the of TL using one triplet. The arrows of each image indicate the direction in which the embedding will move following the gradient.

to optimize the loss. As a consequence, a sampling strategy for the images becomes an essential part of the learning method. The right choice of triplets can increase the speed of convergence, the probability to find a lower minimum and even the possibility of getting better generalization.

In the literature, we can find two different methodologies to face the problem of triplet sampling for each batch. The first one is called *Batch All*, being introduced in Ding et al. (2015), TL_{BA}. In this case, for each sample x_a in the batch, we consider all possible triplets. This results in $k_0 \cdot k_1 \cdot (k_0 + k_1 - 2)$ elements. The TL_{BA} loss function is:

$$L_{BA}(\tau) = \sum_{c=1}^{2} \sum_{a=1}^{k_0} \sum_{\substack{p=1 \\ p \neq a}}^{k_0} \sum_{n=1}^{k_1} \Big[\|f^a - f^p\|_2^2 - \|f^a - f^n\|_2^2 + \alpha \Big]_+ \tag{4.5}$$

The use of the previous methodology declined from the appearance of the *Batch Hard* (Ding et al., 2015) approach, TL_{BH}. It takes each anchor x_a and generates triplets by seeking in the batch for the hardest positive sample x_p, defined as the farthest positive sample $x_p = argmax_{x_i}(\|f_i^a - f_i^p\|_2^2)$, and the hardest negative sample x_n, defined as the closest negative sample $x_n = argmin_{x_i}(\|f_i^a - f_i^n\|_2^2)$. The TL_{BH} loss function is:

$$L_{BH}(\tau) = \sum_{a=1}^{k_0+k_1} \Big[\underset{x^p}{argmax}(\|f^a - f^p\|_2^2) - \underset{x^n}{argmin}(\|f^a - f^n\|_2^2) + \alpha \Big]_+ \tag{4.6}$$

4.3.3 Evaluation

In the field of medical imaging, and in particular, when databases are protected and not released to the public domain, the evaluation of different proposals is perhaps the hardest and most critical part. However, as shown in the related work, a unified procedure that

allows an objective comparison of methods does not exist. We can see that a diversity of metrics are used and in most of the cases, not the ideal ones for the problem. Moreover, in most of the papers, the used or the reported information about the dataset is not sufficient to understand the relevance of the proposal. To this aim, in this section, we study and propose a methodology to be used in order to validate computer-aided polyp detection systems. The proposed evaluation methodology is divided into three fundamental points:

- Databases and cross-validation strategy: How to properly build it and what information must be reported to understand the relevance of the proposal and allow a detailed comparison of models.
- Quantitative Results: Standard metrics in computer vision problems have several drawbacks that can affect the understating of the performance of the methods. For this reason, we propose and justify a set of metrics to be used.
- Qualitative Results: Aside from the numeric results, it is important to consider qualitative results to trust in the system. To this aim, we propose the use of a method to understand the output of the model.

Databases and cross-validation strategy

The creation of medical databases is an essential step before training and validating any type of system. Both, positive and negative samples must be collected in the best way. With respect to size and diversity, training data can follow any distribution that one deems appropriate, however, it is crucial that results are reported using a test set large enough to also capture the diversity of non-pathological images that can be found. In order to capture this diversity, we consider a uniform time sampling strategy as the best option for creating the negative set. As negative samples are cheap, since we have as many as needed, a minimum number of images per video should be used, being 2,000 a reasonable number.

The second important point to consider when creating the database and its evaluation methodology is that although all polyps have common visual characteristics, the appearance of different polyps from the same patient must be considered. The first row of Figure 4.5 shows three different polyps from the same patient, while the second row shows three polyps from different patients. As it can be observed, those polyps from the same patient are generally similar in shape and texture while the polyps from other patients are more diverse. It is for this reason, that training and test sets must not use images from the same procedures.

Additionally, since the datasets are small, it is recommended to perform cross-validation to avoid data selection problems. As mentioned before, it is important that the folds of the cross-validation process are done by leaving procedures out, not by leaving images out, therefore ensuring that images from the same procedure never belong to two different partitions.

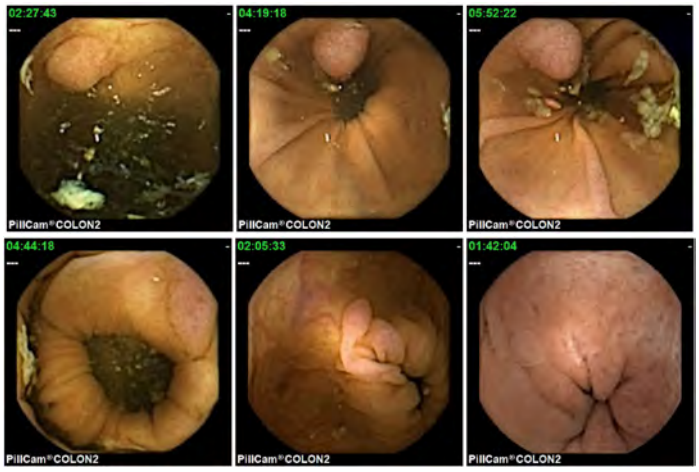


Figure 4.5: Example of polyps extracted from the procedures. In the first row, the polyps come from the same procedures, while the polyps from the second row come from different ones.

Lastly, and since in most of the cases databases are not released to the public domain, it is fundamental to have a detailed description of the dataset in order to understand the complexity and impact of the solution. We consider that the following information should necessarily be reported:

- Number of procedures/cases used in the dataset.
- How many of them suffer a pathology?
- Distribution of unique pathological events.
- Frames per each pathological event.

Quantitative Results

Evaluation metrics illustrate the performance of the system and allow to compare. For this reason, they require a high capability of discrimination among models and they must match with the aim of the system.

Accuracy is the most frequently used metric to validate polyp detection systems. However, it does not reflect what is expected from the system since it does not necessarily correlate with the time needed to reach a diagnosis by the physician. Accuracy depends on the defined threshold of the system, without giving the full picture of the system output. Moreover, in imbalanced problems, accuracy is mostly affected by the predominant class. Weighted accuracy is a more suitable metric, although it is still dependent on a fixed threshold.

Precision and recall (also known as sensitivity), have also been widely used by the

community. Precision is the fraction of true polyp images (TP) among all the positives obtained by the system (TP + FP), while recall is the fraction of true polyp images (TP) that have been detected from the total amount of polyp images (TP + FN). However, with accuracy, these measurements are also affected by the threshold of the classifier. Since the goal of the system is to reduce the time needed for the physician, it is interesting to report the obtained recall scores at different specificity (TNR) values instead of using the best trade-off between both metrics. The recall at these fixed specificity values allows us to understand the expected amount of images that are needed to be visualized by the physician in order to obtain a certain performance, i.e., recall at the specificity of 95% measures the percentage of detected polyps (TPR) if only 5% of the images are reviewed. The recall for specificity values of 80%, 90% and 95% are analyzed for this paper.

The Area Under the ROC Curve (AUC) is another interesting measurement. The AUC is computed from the ROC curve which relates the specificity and the recall obtained for all the possible thresholds of the classifier. The AUC value can be understood as the probability of the classifier to predict a true positive element as a positive with higher probability than as a negative; therefore the larger the AUC value is, the better the classifier is. A limitation of the AUC is that both negative and positive classes have the same impact on the output, so FP and FN penalize equally.

Qualitative Results

Although DL has shown impressive results, its application to the medical field carries worries and criticisms since computerized diagnostic methods are seen as black boxes which do not show how the data is analyzed or how the output is obtained. In medical imaging problems, and particularly on the topic of polyp detection, it is transcendent to trust and understand the obtained predictions by the system. Understanding how the outcome was obtained allows to: 1) understand why something is considered pathological; 2) provide the needed trust of physicians and scientists in the system; 3) debug and identify errors of the system in an easier way.

To this aim, we consider that a qualitative evaluation, showing where and why the system is failing is a very important element. It is not the same to fail in a small and or partly-occluded polyp than missing a large polyp. It is also important to show FP cases, since these images with shapes or textures that are similar to polyps may be understandable errors and increase the confidence in the system.

Studying where the system detected a polyp in a frame can be useful for two main reasons: 1) to identify if the detector is focusing on the area of interest and 2) to help physicians in the review process.

Class Activation Map (CAM) presented by Zhou et al. (2016), is a generic localized deep representation that could be used to interpret the prediction decision made by the system. This method indicates the implicit attention that the network gives to the pixels

of an image considering the class where it belongs.

To obtain the class activation map a linear combination is computed between the feature maps and the classifier weights, since they connect the output and the last feature maps, which identify the importance of each response obtained. For a given class c, the formalization of the class activation map M_c is defined as:

$$M_c(x, y) = \sum_k w_k^c \cdot f_k(x, y) \tag{4.7}$$

where K is the number of channels, f_k are the different responses of the filters and w_k^c is the weight that relates class c with filter k, which is activated by some visual pattern within its receptive field. Briefly, this technique is a weighted linear sum of these visual patterns at different spatial locations, which gives the most relevant regions for the system.

4.3.4 Guidelines

After analyzing and discussing each one of the previous aspects, in order to get a full validation of the system, we propose that the following items should be included in the validation methodology:

1. A fully detailed report of the dataset used.
2. Training and validation set must not contain images from the same procedure.
3. The negative images in the validation set must represent the diversity of the domain. A random sampling or uniform time sampling from the same videos are good strategies (patients and control cases).
4. The number of negative images in the validation set must be higher than the number of positive images. At least 2,000 times the number of positive images should be considered.
5. For small datasets it is necessary to apply a cross-validation method.
6. Recall@80, Recall@90 and Recall@95 should be used in order to make the system comparable with other methods in the community.
7. A qualitative evaluation is recommended to build trust in the system.

4.4 Experimental Setup and Results

4.4.1 Dataset Details

The database used in this paper is composed of 120 procedures from different patients. All these procedures have been performed using Medtronic PillCam COLON2 or PillCam SB3.

Each image from the video was labeled as positive, where at least one polyp was visible, or negative. All these labels were obtained by expert physicians and trained nurses. Each video was examined by at least two experts. In case of controversy between experts, the final decision was taken by a final expert. Polyps were found in 52 out of the 120 analyzed procedures. From those 52 procedures with polyps, a total of 165 different polyps were annotated and used as a positive set. Table 4.3 summarizes the number of polyps found per procedure. As it can be observed, the number of polyps per procedure is diverse, in the majority of procedures, the experts have not reported any polyps, being 1.37 the average of reported polyps per procedure and 11 is the maximum number in a single procedure. Table 4.4 shows the number of frames where each polyp is visualized within the video. Since most polyps are observed in more than one frame, a total of $2,136$ images with polyps have been considered as positive images. Additional details of the database are reported in Table 4.5, such as the morphology and size of the polyps. The size of the polyps was determined using the Rapid PillCam Software V9.

Figure 4.6 shows 9 polyp samples of different sizes and morphologies.

All the images have 256×256 resolution and the time stamp and device information were removed.

Table 4.3: Amount of polyps per procedure.

# Polyps	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**11**
# Procedures	68	17	11	8	3	5	3	2	3

Table 4.4: Amount of frames per polyp.

# Frames	**1-2**	**3-4**	**5-6**	**7-10**	**11-20**	**21+**
# Polyps	33	32	20	19	31	30

Table 4.5: Morphology - Size of the polyps

		Morphology			
		Sessile	**Pedunculated**	**Undefined**	**Total**
Size	**Small** (2-6 mm)	65	4	19	88
	Medium (7-11 mm)	29	4	20	53
	Large (12+ mm)	8	3	13	24
	Total	102	11	52	165

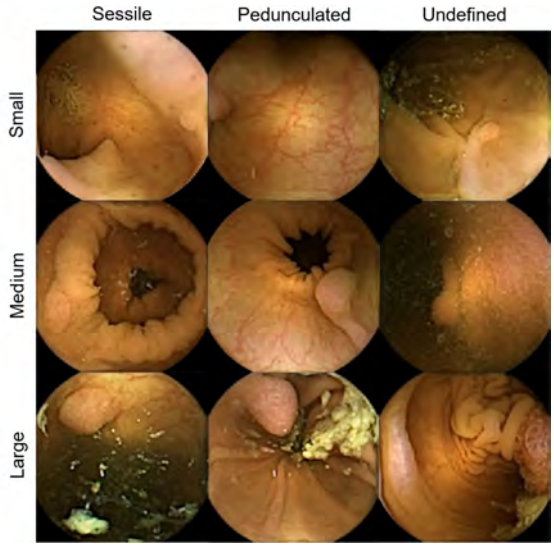


Figure 4.6: Polyp samples of different sizes and morphologies. Columns represent the different morphologies while raw represents the different sizes of the polyps.

4.4.2 Architecture and Evaluation Details

In all of the experiments, a pre-trained model with the ImageNet dataset is used to alleviate the problem of data scarcity. Moreover, in order to enlarge the number of available images, data augmentation for training is performed by applying rotations of $0, 90, 180, 270$ degrees, horizontal and vertical flips, and changes in the brightness of the images with a random probability.

Networks were optimized using Stochastic Gradient Descent with a fixed learning rate of 10^{-3} during 50 epochs. The hyper-parameter margin of the TL was fixed at 0.2. The batch size was fixed to 64, and the proportion of positive and negative images was set to $1/5$ and $4/5$ respectively. In order to not create a bias on the large videos, negative images were obtained using stratified random sampling from those procedures in the training set. Since the dataset is highly imbalanced, an epoch is considered once the entire set of positive images is passed forward and backward through the neural network.

To assess performance, results are reported following the 5-fold cross-validation strategy. It is important to remark that the stratified partitions have been done not by individual frames but by patients, thus images from the same patient do not belong to the same partition of the validation set.

The implementation of the methods has been done using TensorFlow and executed on a machine with an NVIDIA GeForce RTX 2080 TI. Training the network for 50 epochs takes about one hour, and the processing time per image in a forward pass is approximately $2.7ms$. Taking into consideration that the mean number of frames per procedure in our database is $15,183$, the mean processing time per video is $40.89s$.

4.4.3 Quantitative Results

In the first experiment, we aim to compare the performance of each one of the methodologies explained previously: *ResNet*, TL_{BH} and TL_{BA}. As shown in Table 4.6, our methodology TL_{BA} has outperformed the obtained results by the standard optimization methodology of *ResNet* and TL_{BH}. Contrary to what usually happens, batch all sampling strategy exceeds the batch hard strategy. To our understanding, this result is due to the complexity of the generated triplets. Images from the same class are not visually similar. On one hand, all (positive and negative) images present a high visual variability due to the free movement of the camera, the different parts of the gut, or because of intestinal content, such as food in digestion or bile, and on the other hand, polyps can be found in different stages presenting different sizes and morphologies. The obtained AUC value of our methodology has a 12% increase with respect to *ResNet* and a 5% compared to TL_{BH}, achieving $92.94 \pm 1.87\%$. The system enhancement is also reflected in the obtained sensitivity scores, which increased between 25 and 40 points compared to the other models. This fact shows that, given the same percentage of frames reviewed by the experts, the system finds more pathological frames. Figure 4.7 shows the ROC curves of the three studied models. On the left side of the curves, the TL_{BA} model obtains a higher recall value than the other methods considering the same specificity. This difference means that TL_{BA} detects more frames containing polyps, while at the right side of the curve the three systems, TL_{BA}, TL_{BH} and *ResNet*, work similarly. It is remarkable to notice that in the TL_{BA} experiment, the low standard deviation values indicate that the model is more robust than the others.

The proposed method is also compared against some of the most recent polyp recognition systems: 1) *SSAEIM* from Yuan and Meng (2017), 2) *UDCS* from Yuan et al. (2020) and 3) *ANET* from Guo and Yuan (2019). All these methods have been implemented, trained and evaluated using the same dataset and evaluation methodology. The full details of the results are shown in Table 4.6 and Figure 4.7. As it can be observed, the proposed TL_{BA} model demonstrates a significant improvement over these state-of-the-art systems. The system shows an increment of at least the 2% in the AUC value and an increase of around 4% in the different sensitivity values. On one hand, we observed that the *SSAEIM* method is not able to handle this imbalanced and complex database. From our point of view, as a consequence of the complexity of the problem, the dense layers of the autoencoders were not able to find a proper representation of the images. On the other hand, and although *UDCS* and *ANET* show good results, the obtained AUC score and the sensitivity values do not reach the results obtained with our method. This result reflects the capacity of our method to detect polyps, validating the efficiency of TL in this application.

In the next experiment, see Table 4.7, we contrast the performance of TL_{BA} against the same methodology but adding a new extra layer, that acts as the embedding layer. This experiment is done since when deep metric losses are used, it is common to add an extra dense layer between the extracted features and the classification layer. This layer introduces more versatility in data representation while it compresses the information in

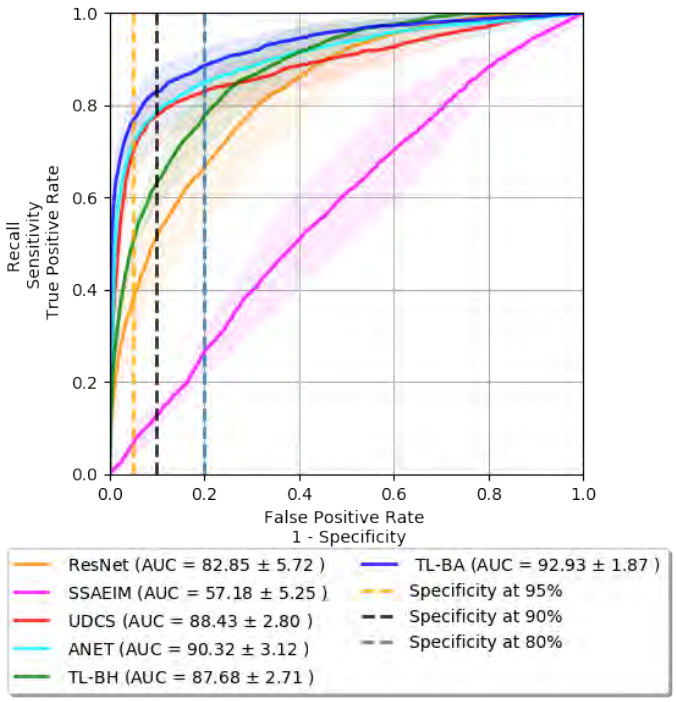


Figure 4.7: ROC Curve of the three models. Each vertical line represents a specificity value that indicates the percentage of true negative images predicted in the video, and the percentage of polyps that the system is expected to detect.

the embedding. As shown in Table 4.7, the embedding sizes used in these experiments are: 128, 256, 512 and $1,024$. Despite the fact that the new networks have more parameters, none of them exceeds the previous TL_{BA} results in the AUC score. The obtained AUC value and sensitivity values show a correlation between the embedding size and the obtained scores. TL_{BA} model with an embedding size of 2,048 has exceeded the other models because the variation in the embedding size allows the network to have a better representation to detect the polyps.

The margin hyper-parameter of the TL has been set until now at 0.2 as it is set in other works like Schroff et al. (2015) or Hermans et al. (2017). As the domain of the problem is different from previous applications of the TL method, our fourth experiment evaluates the behavior of the system with the following margins: 0.1, 0.5 and 1.0. As shown in the obtained results summarized in Table 4.8, any of these margins outperforms all the metrics. Margins 0.5 and 1.0 obtain standard deviation values which are higher than the small margins, indicating that the model is less robust. However, the margin that achieves the best results on almost all the reported metrics is 0.2.

The comparison of models is shown in Tables 4.6, 4.7 and 4.8 demonstrating that TL_{BA} is the best computer-aided decision support system for polyp detection in terms of accuracy and sensitivity.

Table 4.6: Performance comparison of the methods: ResNet, TL_{BA} and TL_{BH}. Each method has been evaluated with a 5-fold validation in the classification task.

Parameter Optimization	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Sensitivity (%) Spec. at 95%	Spec. at 90%	Spec. at 80%
ResNet	97.85 ± 0.24	26.01 ± 9.78	97.97 ± 0.27	82.85 ± 5.72	37.75 ± 9.12	51.49 ± 11.09	66.71 ± 12.15
SSAEIM	59.85 ± 48.75	40.11 ± 48.90	59.91 ± 48.92	57.76 ± 5.83	6.98 ± 2.99	13.29 ± 3.56	27.82 ± 5.96
UDCS	94.41 ± 1.53	71.51 ± 7.80	94.45 ± 1.54	88.64 ± 2.87	70.44 ± 6.53	78.22 ± 6.46	83.31 ± 5.18
ANET	96.96 ± 0.53	$\mathbf{65.07 \pm 7.58}$	97.02 ± 0.54	90.44 ± 3.23	72.02 ± 6.03	78.92 ± 5.59	85.23 ± 4.98
TL_{BH}	$\mathbf{99.83 \pm 0.05}$	0.00 ± 0.00	$\mathbf{100.00 \pm 0.00}$	87.68 ± 2.71	50.15 ± 3.21	63.19 ± 4.48	77.52 ± 6.70
TL_{BA}	$\mathbf{99.43 \pm 0.12}$	51.15 ± 7.62	99.51 ± 0.12	$\mathbf{92.94 \pm 1.87}$	$\mathbf{76.68 \pm 4.93}$	$\mathbf{82.86 \pm 4.78}$	$\mathbf{88.53 \pm 3.76}$

Table 4.7: Performance of the methods: TL_{BA} and different versions of the same adding an extra dense layer and changing the embedding size. Each method has been evaluated with a 5-fold validation in the classification task.

Embedding	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Sensitivity (%) Spec. at 95%	Spec. at 90%	Spec. at 80%
TL_{BA}	$\mathbf{99.43 \pm 0.12}$	51.15 ± 7.62	$\mathbf{99.51 \pm 0.12}$	$\mathbf{92.94 \pm 1.87}$	$\mathbf{76.68 \pm 4.93}$	$\mathbf{82.86 \pm 4.78}$	$\mathbf{88.53 \pm 3.76}$
128	59.99 ± 7.48	$\mathbf{84.67 \pm 6.17}$	59.95 ± 7.51	83.05 ± 2.78	42.23 ± 5.04	57.38 ± 5.54	72.54 ± 4.85
256	68.12 ± 12.92	79.10 ± 14.60	68.09 ± 12.97	83.35 ± 5.27	45.09 ± 10.34	59.9 ± 11.64	73.35 ± 10.11
512	86.29 ± 2.56	70.50 ± 9.02	86.31 ± 2.57	86.58 ± 3.91	49.91 ± 8.56	64.66 ± 6.43	78.13 ± 6.04
1024	89.69 ± 2.50	68.28 ± 9.44	89.72 ± 2.51	86.67 ± 3.25	55.47 ± 3.86	67.97 ± 5.79	79.89 ± 5.29

From a medical point of view, the computer-aided system should help to detect polyps but not necessarily detect all the images where a given polyp is seen. For this reason, we analyzed the performance of our proposed system over polyps. A global overview of the numerical results is summarized in Table 4.9, where each score represents the percentage of detected polyps in different scenarios of the entire dataset. Each of them is computed with a different specificity value: 80%, 90% and 95%. The first row of the table contains the percentage of detected polyps, that grows when the specificity decreases. Setting the specificity at 95%, the system only misses 14 polyps; if we decrease specificity to 90% and 80%, the missed polyps are 11 and 8 respectively. A complete view of the curve is reported in Figure 4.8.

The second set of results in Table 4.9 presents the detection of the system according to polyp size. When we consider small polyps the amount of missed polyps is 7, 4 and 3 for the respectively reported values of specificity. In the case of medium-sized polyps, using the specificity of 90 or higher, the system is not able to detect 5 polyps, but with lower specificity, the amount of detected polyps raises. In the case of larger polyps, 2 of them are lost for the two higher specificity values, however, with a slight decrease in specificity, the polyps are detected.

Finally, the last rows of Table 4.9 show the detection rated based on polyp morphology: sessile, pedunculated or undefined. As it has been reported previously, most polyps are labeled as sessile, obtaining high detection scores despite the misplacement of 7. Pedunculated polyps are relatively rare, and the system detected all except one at the three

Table 4.8: Performance of TL_{BA} method changing the margin parameter. Each network has been evaluated with a 5-fold validation in the classification task.

Margin	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Sensitivity (%) Spec. at 95%	Spec. at 90%	Spec. at 80%
0.1	87.98 ± 2.01	66.85 ± 7.75	88.01 ± 2.01	86.53 ± 2.96	47.35 ± 10.22	63.35 ± 7.66	78.69 ± 5.60
0.2 (TL_{BA})	$\mathbf{99.43 \pm 0.12}$	51.15 ± 7.62	$\mathbf{99.51 \pm 0.12}$	$\mathbf{92.94 \pm 1.87}$	$\mathbf{76.68 \pm 4.93}$	$\mathbf{82.86 \pm 4.78}$	$\mathbf{88.53 \pm 3.76}$
0.5	90.89 ± 2.07	$\mathbf{68.79 \pm 8.66}$	90.93 ± 2.07	87.74 ± 3.04	56.07 ± 9.18	70.96 ± 7.44	83.87 ± 4.61
1.0	91.67 ± 1.00	66.30 ± 8.74	91.71 ± 1.02	86.80 ± 3.23	56.84 ± 8.68	70.65 ± 7.38	81.02 ± 6.06

sensitivity values.

Table 4.9: Detection vs. Specificity with model TL_{BA}

% ***detection***	***Specificity@95***	***Specificity@90***	***Specificity@80***
Polyps	91.14%	93.04%	94.94%
Small Polyps	91.86%	95.35%	96.51%
Medium Polyps	90.38%	90.38%	92.31%
Large Polyps	90.00%	90.00%	95.00%
Sessile Polyps	93.07%	95.05%	96.04%
Pedunculated Polyps	90.91%	90.91%	100.00%
Undefined Morphology	86.96%	89.13%	91.30%

4.4.4 Qualitative Results and Polyp Localization

CAM visualization was applied to the output of the network. This method generates a heat map, where the red tones show the regions of the image that obtain a high response from the filters. Figure 4.9 shows in the first row eight polyps frames where the different morphology and size of the polyps may be observed. In the second row, the CAM visualization method highlights the location where the system focused to predict that there was a polyp.

Figure 4.10 shows eight images without polyps where the system has erroneously detected a polyp. In these samples, some of the regions highlighted by the network contain features of polyps such as growths of tissue, mucous membranes or areas with reddish color from the wall, that might indicate the existence of it.

Figure 4.11 shows eight polyp images where the system has not obtained enough features to predict the frame as a polyp. Each image shows a boundary with the location of the polyps. These difficult cases are complex to detect in single images by the system. The evaluation of a whole sequence of images where the polyp is seen facilitates detection by the human eye. Due to the complexity of polyp detection, sometimes is easier for humans to detect them through the sequence.

Figure 4.12 shows the second sequence of images in Figure 4.1 with the output of the system represented by adding a green square around the frames where the system has detected a polyp. Although in this example the system missed two frames where the

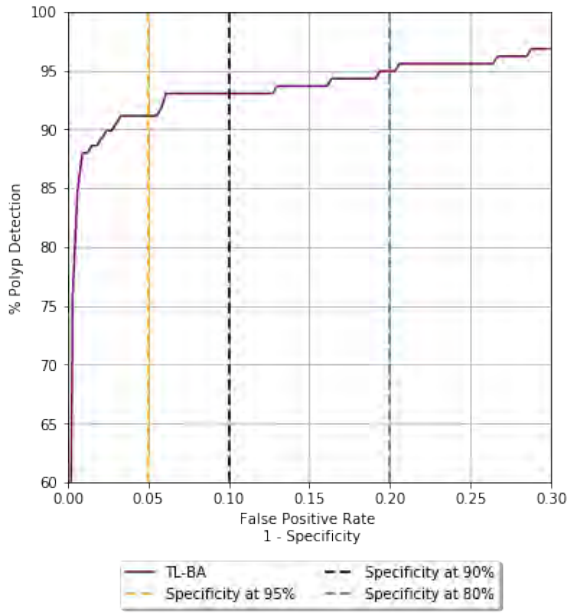


Figure 4.8: Percentage of polyps detected.

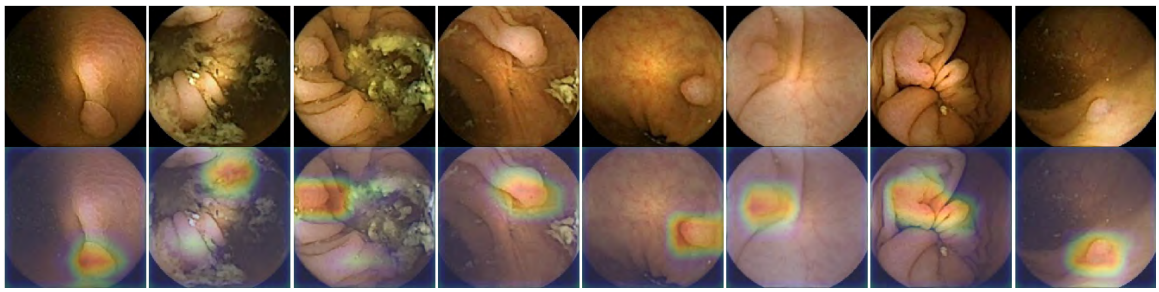


Figure 4.9: The first row contains eight examples of TP of our proposed method, with polyps of different morphology and size. The second row incorporates the CAM representation that locates each one of the polyps over the original image.

polyp is present, the detection in four frames is sufficient for the physician to establish the diagnosis.

4.4.5 Effect of imbalance datasets over models

Healthcare datasets commonly suffer from imbalanced data, a feature that frequently affects the performance of classical DL approaches. In the following experiment, we evaluate how both, TL and the original ResNet networks behave with respect to different degrees of imbalanced data. The imbalance degree of a dataset is defined as the number of negative images per each positive one. To conduct a meaningful comparison, both models are trained and evaluated on six different imbalance degrees (1, 10, 25, 50, 100). Furthermore, each imbalance degree is repeated ten times using different sampled data.

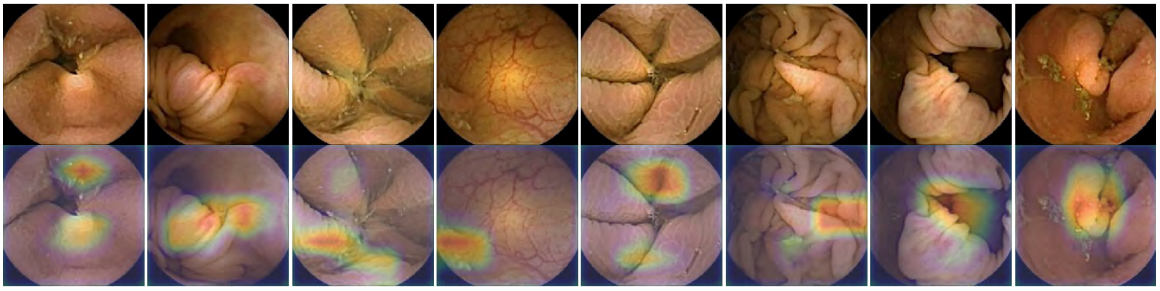


Figure 4.10: The first row contains eight examples of FP of our proposed method, where the system has detected polyps. In some images abnormal tissue can be seen, some mucous membranes or reddish zone, that are features related to polyps. The second row shows the CAM representation that locates where these features are located.

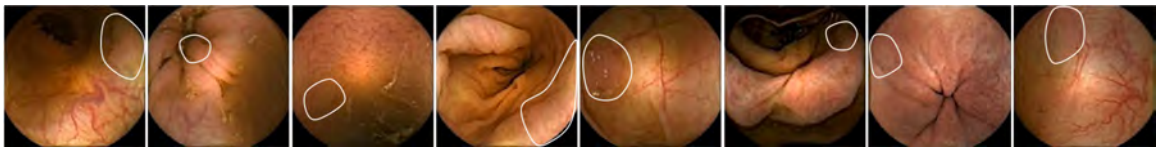


Figure 4.11: The images correspond to eight examples of FN of our proposed method, where a polyp is in the frame, but the system couldn't detect it. To help the reader to find the polyps in the images, the outline of the polyp has been drawn in white color.

As it can be seen in Figure 4.13, the experiment shows the robustness of the proposed methodology to highly imbalanced data. The performance of the method increases when the degree of imbalance increased from 1 to 10 and then stabilizes. On the other hand, it can be observed that the original ResNet network suffers from highly imbalanced data. The network achieves the best result with an imbalance degree of 10 and then it starts to deteriorate as the imbalance degree increases.

4.5 Conclusion

The methodology proposed in this study improves automatic polyp detection in WCE images and additionally enables localization of the polyp in each image. The reported experiments demonstrate that the TL method improves feature extraction outperforming previous results and that the limited and imbalanced data availability may be alleviated with the appropriate losses. Furthermore, the qualitative output of the system may increase trust in the prediction.

Future research will focus on the detection of other intestinal pathologies to develop a complete computer-aided detection system for WCE videos.

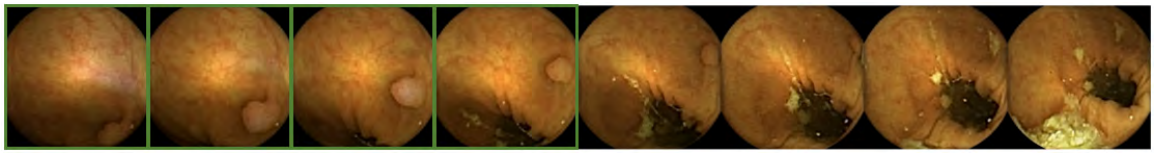


Figure 4.12: Polyp sequence where the green squares denote the presence of polyps detected by the system. In this sequence, there are two frames where the polyp is not detected, despite this, the support system has found the polyp in the previous frames, allowing the doctor to diagnose the patient.

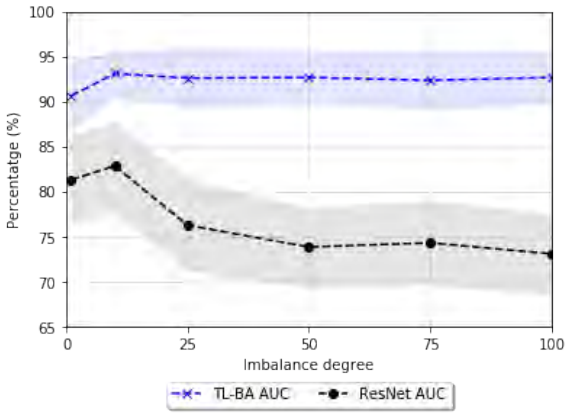


Figure 4.13: AUC values of ResNet and our methodology trained with difference imbalance degrees. Each point represents the mean of 10 executions of a 5-fold validation in the classification task.

Acknowledgment

The authors would like to thank the team from CorporateHealth International ApS for their feedback and economic support and NVIDIA for their GPU donations. This work has been also supported by MINECO Grant RTI2018-095232-B-C21 and SGR 1742 as well as the Innovate UK project 104633.

Chapter 5

Paper II

Time-based self-supervised learning for Wireless Capsule Endoscopy

Guillem Pascual[a], Pablo Laiz[a], Albert García[a], Hagen Wenzek[b], Jordi Vitrià[a], and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

In Computers in Biology and Medicine, Elsevier, Volume 146:105631, 2022. ISSN 0010-482

doi: 10.1016/j.compbiomed.2022.105631

Impact Factor Journal: 6.698
Q1 Computer Science, Interdisciplinary Applications (24/112)

Contents

Motivation and Context

DL models require a large number of samples to be trained and learn the patterns of the data. However, the medical field datasets often contain only a few hundred or thousand examples, making it challenging to train robust models. Additionally, medical datasets also suffer from imbalanced class distributions, where one or more classes have a disproportionately low number of examples compared to the others. This imbalance can lead to models that are biased towards the more represented classes, resulting in poor performance for the minority classes.

To address the previous issues, researchers have adopted transfer learning strategies. They have been widely used in the computer vision community, where pretrained models on large datasets, like ImageNet, are used as a starting point for many image recognition tasks (Liu et al., 2021). However, medical imaging datasets, and in particular, images captured with WCE videos, have different features compared to images containing real-world objects, making it necessary to find better pretraining settings.

To handle the aforementioned problem, this study proposes the use of SSL to extract richer representations from unlabeled data. Instead of using data augmentation techniques and NT-Xent loss, as other SSL methods such as Chen et al. (2020a), the proposed learning approach takes advantage of the temporal axis of WCE videos and uses the TL to optimize

the weights of the neural network. This combination leverages the temporal information present in hours-long capsule endoscopy videos to learn more generalized and rich embeddings through the exploration of inter-sequence and inter-video triplets.

After obtaining the pretrained model using SSL, the network is fine-tuned with the labeled data for two specific tasks: polyp detection and GI lesion classification. The experimental results show that utilizing the structure learned from the temporal axis, as inferred by the proposed model, enhances the detection rate in various domain-specific applications. Furthermore, the detection rate remains effective even when confronted with significantly imbalanced datasets. This highlights the potential of SSL as a promising approach to mitigate the challenges posed by small medical imaging datasets and imbalanced class distributions. Moreover, it opens up new possibilities for improving the performance of DL models for other medical applications.

Time-based Self-supervised Learning for Wireless Capsule Endoscopy

Guillem Pascual[a], Pablo Laiz[a], Albert García[a], Hagen Wenzek[b], Jordi Vitrià[a], and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

5.1 Abstract

State-of-the-art machine learning models, and especially deep learning ones, are significantly data-hungry; they require vast amounts of manually labeled samples to function correctly. However, in most medical imaging fields, obtaining said data can be challenging. Not only the volume of data is a problem, but also the imbalances within its classes; it is common to have many more images of healthy patients than of those with pathology. Computer-aided diagnostic systems suffer from these issues, usually over-designing their models to perform accurately. This work proposes using self-supervised learning for wireless endoscopy videos by introducing a custom-tailored method that does not initially need labels or appropriate balance. We prove that using the inferred inherent structure learned by our method, extracted from the temporal axis, improves the detection rate on several domain-specific applications even under severe imbalance. State-of-the-art results are achieved in polyp detection, with $95.00 \pm 2.09\%$ Area Under the Curve, and $92.77 \pm 1.20\%$ accuracy in the CAD-CAP dataset.

Key words: capsule endoscopy; deep learning; self-supervised learning; semi-supervised learning.

The paper has been re-typeset to match the thesis style.

5.2 Introduction

Obtaining Gastrointestinal (GI) images has traditionally been an intrusive intervention until the advent of Wireless Capsule Endoscopy (WCE) technology (Iddan et al., 2000). WCE imaging eases the process of securing a continuous stream of images, but at the same time, it introduces its own set of problems.

The videos recorded by the capsule, although usually with a low frame rate, can have a duration of up to 12 hours (Vasilakakis et al., 2019). Unlike traditional methods, it is not a targeted exploration but rather a complete recording as the capsule travels through the entire system. A physician must go over the full length of the video, possibly at multi-image speeds, while looking for any abnormality. Not only do they have to invest considerably more time, but the fatigue and repetitiveness of the task could affect their ability to detect such abnormalities.

Providing a reliable and accurate Computer-Aided Diagnosis (CADx) system capable of selecting the most promising frames would ease the pressure for those professionals, cutting down the time spent on the task while obtaining comparable—if not better—results.

Also of great importance, especially when designing automated systems that rely on images obtained from patients, is to examine the properties of the data. In day-to-day examinations, not all patients have an associated pathology, and the data used in research to train CADx models directly reflects it. In polyp detection, for example, the majority of videos have no polyp present in a video at all. One must also consider that, even in the case that there might be polyps, they would appear only in a small fraction of the frames (Laiz et al., 2020). A polyp might appear in several subsequent frames, perhaps slightly displaced or rotated, but the overall number would be negligible when considering the whole duration of the video.

Combining the difficulty of obtaining said datasets with the amount and distribution of the data itself makes creating accurate and production-ready CADx systems a difficult task. Data is fairly scarce compared to other problems studied in deep learning, and the classes, such as polyp, or non-polyp, suffer significant imbalances. Not to mention that supervised algorithms, which dominate the field, require that all those videos are accurately labeled to function.

Creating better models for the medical field which, ultimately, could be used in CADx, requires sorting out these issues. Techniques like data augmentation and regularization have been used to cope with overfitting and under-generalizing models, but they are hard to train and can obtain sub-par results. As such, it is the aim of this work to produce a method that enables obtaining better WCE models without over-relying on these two approaches. The main motivation being that such models would help reduce the workload that physicians are facing when examining WCE videos while, perhaps even more importantly, not sacrificing any accuracy, be it detecting polyps, bleeding, or any other critical condition.

Thereafter, this work proposes the application of Self-Supervised Learning (SSL) on WCE videos to obtain a better representation of the data, enabling future models to perform better in their classification tasks. Self-supervision has been canonically considered a variant of supervised learning (Liu et al., 2021), as the network learns from supervisory signals obtained from the data itself, often leveraging the underlying structure in the data. Based on this definition, we derive a novel pseudo-labeling method for WCE that works with several unlabeled videos, enabling the use of SSL, which helps train models for downstream tasks.

In SSL, instead of directly training a model with a set objective in mind, the process is divided into two steps. SSL is done during an initial phase named *pretrain*, where a deep neural network is trained to learn a better representation, or embedding, of the data. It encodes the most essential information into a smaller vector by using the data without their final labels, learning its inherent structure. This information is learned accordingly to the data's nature, the model's architecture, and the task used for SSL. Then, during a second pass, the *finetune* process, the embedding is used in conjunction with the labels to perform supervised classification.

With the present work, summarized in Figure 5.1, we aim to use self-supervision to provide more accurate models for domain-specific tasks derived from WCE images. In particular, given unlabeled WCE videos, we exploit their temporal nature to perform SSL and then train several supervised models. These models can then be used for CADx, which would improve the results with respect to current methods, reducing the workload for physicians.

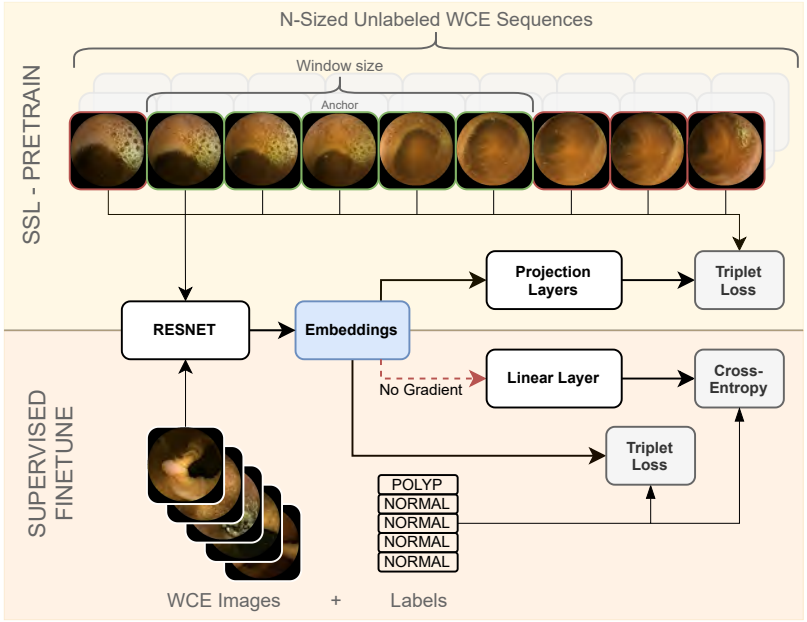


Figure 5.1: Overview of the proposed method, including the pretain phase, in the upper half, and the final finetune phase in the lower half.

The paper is organized as follows. First, we give an overview of the related work in the field followed by a description of our methodology, presenting the self-supervised training, supervised training, and system architecture. Further, we explain the experimental setup and results, and finally present the main conclusions and give directions for future work.

5.3 Related work

5.3.1 Wireless Capsule Endoscopy research

WCE, due to the nature of its long data streams, has been a popular candidate for computer-aided automation. For instance, bleeding detection was first done by means of superpixels in conjunction with a support vector machine (Fu et al., 2014), by means of Super-Vector Machines (SVM) and manually found color invariants (Lv et al., 2011), through saliency maps (Yuan and Meng, 2015), and using hand-crafted textures and multiple machine learning algorithms like classification trees, random forests, and logistic model trees (Pogorelov et al., 2019). Other tasks explored are polyp detection through image subdivision and SVM (Alexandre et al., 2007), ulcer detection with texture and color invariants (Yeh et al., 2014), and motility events with pattern recognition, color decomposition, and chromatic stability (Malagelada et al., 2008).

These processes saw an increase in performance with the advent of deep-learning based models. In Seguí et al. (2016), Convolutional Neural Network (CNN) automate the process of texture finding, no longer requiring hand-crafted features, and achieving better results at motility event classification. Likewise, other WCE domains such as polyp detection (Iakovidis et al., 2018; Aoki et al., 2019; Nadimi et al., 2020), bleeding (Caroppo et al., 2021; Khan et al., 2020), ulcer detection (V and Prashanth, 2020), and celiac disease diagnosis (Wang et al., 2020), have benefited from the use of CNNs.

Recently, WCE models have thrived with more advanced methods, as the works in Laiz et al. (2020); Jain et al. (2021); Yuan et al. (2020); Kundu and Fattah (2019); Jain et al. (2020); Guo et al. (2022) demonstrate. Attention mechanisms to let the network learn the important features (Jain et al., 2021), the use of residual connections with the ResNet model (He et al., 2016) along with metric learning with Triplet Loss (TL) (Schultz and Joachims, 2004) in Laiz et al. (2020), and the ability to created deeper and denser models (Yuan et al., 2020) have enabled them to produce more robust and accurate methods. Noteworthy, disease detection in the gastrointestinal tract has greatly gained from recent advances, with CADx systems being explored in bleeding detection, vascular lesions, ulcers, polyp, and tumors (Trasolini and Byrne, 2021; Attallah and Sharkas, 2021; Gilabert et al., 2022).

Notwithstanding the recent advances, both traditional machine-learning based methods and the deep-learning variants, suffer from the same problems—lack of labeled data and, in some domain-specific tasks like polyp detection, also highly imbalanced classes. This

is formalized and analyzed in Yuan et al. (2020) and Akay and Hess (2019), where the difficulties of producing models that generalize and do not overfit, product of imbalance, low inter-class variance, and high intra-class variance are inspected in detail. Techniques like dropout, L1 or L2 regularization, and sampling mechanisms have also been applied to WCE in an attempt to soften the problems derived from data imbalances and inter-class and intra-class variances, such as overfitting and failing to generalize (Kim and Lim, 2021). Other works, like Laiz et al. (2020), show that using TL to learn better embeddings also contribute towards obtaining more robust models. Nonetheless, the problem still remains, WCE is tedious to label due to its length, which often means that researchers have low amounts of labeled data to work with. Moreover, several fields must still tackle with huge imbalances within the data.

5.3.2 Self-supervised learning

Other approaches to tackle low amounts of labeled data, when pseudo-labeled data is available, and class-imbalances in downstream tasks are self-supervision methods, of which a wide range of options are available. For instance, a popular architecture choice was autoencoders (Rumelhart et al., 1986; Kingma and Welling, 2014; Hinton et al., 2011), whose dimensionality-reducing capabilities were believed to be useful for SSL. However, it has been demonstrated that they fail to capture rich information (Bengio, 2009), focusing only on compressing data. Thus, their capacity to adapt to any future generic task is hindered at best.

In contrast to the former generative method, where the network learns from a single image, contrastive learning trains on multiple examples or instances of the same image to learn the inherent information (Falcon and Cho, 2020). One such way to introduce multiple samples of a single image has been by reordering subsections (Misra and van der Maaten, 2020). This type of SSL encourages the network to learn invariant representations, unlike their generative counterparts. Similarly, when the time dimension is available, reordering can be done based on fragments of the input, as done with audio streams (van den Oord et al., 2018). Additional techniques, like rotation, color jittering, blurring, and cropping, can be applied as shown in Chen et al. (2020a) and Chen et al. (2020b). The authors propose SimCLR, an architecture based on ResNet (He et al., 2016) that can be trained with multiple contrastive approaches and a new contrastive loss. They provide a simple framework to perform SSL and benchmark the different methods.

More specifically and related to our application, contrastive SSL from videos has been done by predicting the order of a sequence (Misra et al., 2016; Xu et al., 2019; Lee et al., 2017), object tracking (Pathak et al., 2017; Wang and Gupta, 2015; Wang et al., 2019d), and specialized losses (Tschannen et al., 2015; Sermanet et al., 2018). In particular, our method resembles the single-view approach of Time-Contrastive Networks (Sermanet et al., 2018), which uses metric learning for temporal coherence. Their work, however, diverges from ours

because they do not focus on the embeddings' richness nor task-generalization. Given the nature of their action imitation task, they limit their triplets to be in a single sequence and do not explore the embedding quality, whereas our work aims to learn generalized and rich embeddings from hours-long videos, exploring inter-sequence and inter-video triplets, for further usage in downstream tasks.

In medical imagining, some efforts have been made in regards to SSL and semi-supervised training (Cheplygina et al., 2019; Azizi et al., 2021). For instance, Pérez-García et al. (2021) uses a generative network to create simulated postoperative MRI images, which used in a SSL step obtains better results. Other tasks such as pneumonia detection and multi-organ segmentation (Navarro et al., 2021), also show improvements by means of SSL based on samples' patch reordering. Likewise, SSL has also been applied to WCE related tasks, using distortions to the original images in Vats et al. (2021), combined with multi-task learning to detect inflammatory and vascular lesions. Similarly, Guo and Yuan (2020) minimizes the difference in predictions between the SSL head and the supervised head, leveraging unlabelled data.

To the best of our knowledge, however, there has been no work that leverages the temporal aspect of WCE videos, using a SSL process to obtain better representations, which, in turn, would help tackle data-derived problems.

5.4 Method

An overview of our proposed self-supervised approach is illustrated in Fig. 5.1. Similar to most methods relying on self-supervised training, our approach is divided into two distinct stages: (a) pretraining a self-supervised network using unlabeled data to obtain rich representations, and (b) finetuning the model using labeled data for a specific task. This section follows the same pattern, explaining both phases first, and finishes by explaining the architecture used.

5.4.1 Self-supervised pretraining

During the first stage of the process, we aim to extract useful generic information from the unlabeled images, which then can be transferred to deal with many specific tasks by finetuning the model with limited labeled data. In other words, it creates a reduced representation (embedding) of the original image that contains its most important information.

Extracting an embedding can be understood as a process $f(x)$, where a neural network transforms a sample x from the dataset to its compressed and rich representation.

Out of all the possible ways to obtain said embedding, we have chosen to exploit the temporal nature of WCE videos. Our method works by taking sequences of N contiguous frames and creating a relationship between them. Namely, given two frames i, j in the

sequence, their relationship is established as the distance $d(i, j)$ between them, counted by the number of frames that separates them.

Unlike the work in Sermanet et al. (2018), where all samples come from a single sequence, our method must generalize to multiple videos and sequences. Per-frame pseudo-labels are introduced to encode their video identifier along with their position. Given an image i, its pseudo-label is a combination of its video identifier $\gamma(i)$, which can be a simple numbered sequence, and the position inside the video $\delta(i)$, as seen in Equation 5.1.

$$\bar{y}(i) = M\gamma(i) + \delta(i) \tag{5.1}$$

Where M must be a large enough number so that $\forall i, M > \delta(i)$. For our particular experiments and datasets, we have chosen $M = 10^6$.

Next, we impose a similarity measure between frames on the sequence so that contrastive learning can be done by finding the inherent relationship between similar and dissimilar images. For that purpose, two images will be consider similar if they are close enough, formalized as $d(i, j) = |\bar{y}(i) - \bar{y}(j)| \leq w$, where $w \leq N$ is a constant chosen beforehand. The pair (i, j) is considered similar (positive) in such cases, and negative otherwise.

In other words, taking a reference image (anchor) in a sequence, all other images within a window of size $2w$ (w images per side) are considered similar. In general, given an N-sequence, all images have between $min(N, 2w)$ and w positive samples. Images around the edges of the sequences lose up to half the positives, tending towards the latter, while those on the center have the whole spectrum.

The pseudo-labels guarantee that (i, j) negative pairs are consistent with images coming from different videos, as $\gamma(i) \neq \gamma(j)$, thus $d(i, j) \approx |M\gamma(i) - M\gamma(j)| \geq M > w$. Additionally, for two frames i, j extracted from the same video, the formula reduces to the distance in frames between them, $d(i, j) = |\bar{y}(i) - \bar{y}(j)| = |\delta(i) - \delta(j)|$.

Given the above approach to create a similarity measure, the TL (Schultz and Joachims, 2004), a contrastive loss, is introduced to learn the embeddings. TL works by using triplets of samples, where two of the triplet's elements, the anchor a and the positive p, pertain to the same class. The remaining element, the negative n, is of a different class than a. That is, given the embedding of an anchor $f(a)$, a triplet $(f(a), f(p), f(n))$ is formed so that $y(a) = y(p) \neq y(n)$, where $y(\cdot)$ is the class of a sample.

Using Equation 5.2, TL forces $f(p)$ to be close to $f(a)$ while moving away $f(n)$. It eases the problem by introducing a soft margin α between the positive and negative pairs.

$$TL = max(||f(a) - f(p)||^2 - ||f(a) - f(n)||^2 + \alpha, 0) \tag{5.2}$$

Translated to our domain, a triplet is formed by two similar images and a dissimilar image, so that $d(a, p) \leq w$ and $d(a, n) > w$. As shown, TL is directly applicable to WCE

videos when used in conjunction with the pseudo-labels, forcing close images in a sequence to have similar representations in the embedding space.

It must be noted that this method is bound to have incorrect pairs, as different videos or sequences could contain similar images, regardless of their distance. Also, WCE videos tend to have periods where the capsule moves at a slow rate, producing many similar images in a relatively long interval, or the contrary, moves fast and captures rapidly changing sequences. We estimate those cases to be negligible compared to our dataset's size, being effectively treated as noise during the process.

5.4.2 Supervised learning

During the second phase of our method, the same model is reused to learn a domain-specific task with limited amounts of data. For instance, the rich representations could be used to model motility events, to classify several conditions like bleeding or inflammation, to evaluate keyframes, or to detect polyps, to name a few.

For that purpose, the process starts with the SSL model's parameters, obtaining embeddings produced by the new dataset and feeding them into a classifier. That classifier needs to access the ground truth labels, as it uses a softmax cross-entropy loss to model the problem.

Following SimCLR findings (Chen et al., 2020a), we have confirmed that fixing the weights obtained during SSL is counterproductive. However, unlike SimCLR, which assumes balanced problems, we use the approach proposed by Laiz et al. (2020), where the TL is used to modify the embeddings. As such, the gradient coming from the linear classifier is removed so that it cannot negatively impact the embeddings due to the imbalance. Instead, a TL is imposed on them to facilitate the network to finetune the dataset representations.

However, unlike in the previous step, the TL no longer uses the pseudo-labels created through our method. Triplets are formed by considering the real labels of the images, which are domain-specific and help finetune the embeddings to the particular task. To further reference it and avoid confusion, the term TL_{sup} will be used.

The TL_{sup} is trained in batch all mode, which considers all triplets regardless of their difficulty. No special sampling algorithm is introduced; the only restriction we impose is for a batch to have a proportional representation from all classes. Other than that, data is randomly sampled.

The final loss obtained in this model is the linear combination of both the cross-entropy loss and the triplet loss, as shown in Equation 5.3.

$$L_{sup} = TL_{sup} + L_{crossentropy} \tag{5.3}$$

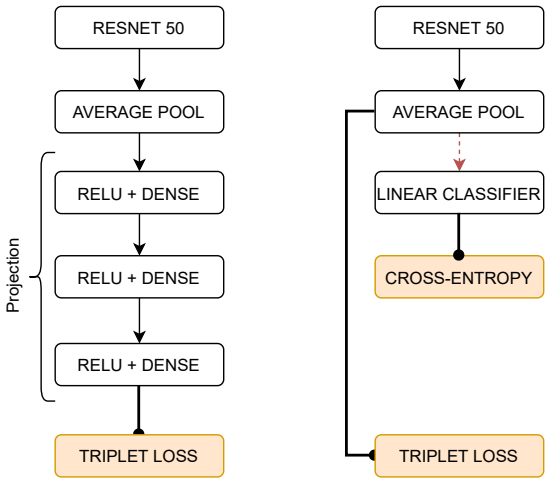
5.4.3 Architecture

The backbone of our architecture consists of a ResNet-50 (He et al., 2016), as can be seen in Figure 5.2a. Most works that extract or require embeddings use the output of the ResNet model directly as their representations, but following the work in SimCLR, we decided to explore the possibility of including several projection layers.

Each projection layer consists of a ReLU activation followed by a dense layer. We restrict all the projection layers to have the same dimensionality, which must be lower than the 2048 given by ResNet. While our pretrain phase benefits from the reduced complexity after the projection, the final finetuning network utilizes the whole 2048-sized embedding to allow for better detection rates. These layers, along with their configuration, hyperparameters, and performance, are studied below. Ultimately, they are found to be beneficial for domain-specific tasks.

Once the pretrain is done, at the beginning of the finetune phase all learned parameters are kept except for the projection layers, which are removed from the model, as can be observed in Figure 5.2b. Classification is done through a linear layer (a dense layer without any activation) and a cross-entropy loss. As denoted in red and a dashed line in Figure 5.2b, we eliminate the gradient coming from the linear classifier to stop it from modifying the embedding. Only the TL loss is able to tune the representations.

It must be remarked that the TL losses used in both phases of the architecture are different. As pointed out, the first phase uses the pseudo-labels deducted from videos, while the second uses the ground truth labels.



(a) Pretrain (b) Finetune

Figure 5.2: Detailed network architecture. The parameters obtained during pretain for ResNet are used in the finetune phase, while the projection layers are removed. Here, the dashed red line denotes that gradient is stopped.

5.5 Discussion and results

This section begins by laying out the datasets used during both steps of the method. Further, it explains the implementation details, such as preprocessing steps and train strategies. A subsection is devoted explicitly to the SSL hyperparameters, justifying and proving the choices made. Finally, individual results are shown for each dataset, discussing the results qualitatively and quantitatively.

5.5.1 Datasets

Three datasets are used throughout the process. The Generic WCE videos dataset is employed only for the common SSL stage, while the other two are each used to evaluate their own downstream tasks.

Generic WCE videos

This dataset consists of a total 49 unlabeled WCE videos, each from different patients, obtained with Medtronic PillCam SB2. From those videos, only the small intestine and colon segments are used, selecting a total of 1,185,033 frames.

Even though these images are not labeled, pseudo-labels can be introduced through our method, which makes this dataset suitable for a pretrain step using SSL.

Polyp WCE

The dataset consists of 248,136 frames sampled from 120 procedures performed using Medtronic PillCam SB3 and PillCam Colon 2. Notably, they are not the same videos as the subsection above. Of those frames, 2,080 contain polyps, while 246,056 do not. An initial report is produced by eight expert readers, endoscopy nurses with at least three months of experience, who tag potential polyp frames, and others that require detailed revision. Then, two medical doctors (one gastroenterologist, one internal medicine) obtain the final version of the dataset. The polyp's sizes, as reported in Table 5.1, were obtained through Rapid PillCam Software V9. The largest polyp was determined to be 16 mm. Tumors were considered positive, while any other pathology, like ileal lymphoid hyperplasia, bleeding, and diverticulitis, were discarded from the dataset.

Unlike the Generic WCE videos, this dataset uses SB3 and Colon 2 as sources. It is shown in Laiz et al. (2019) that using SB3 from SB2 is possible, while Laiz et al. (2020) demonstrates that having mixed sources poses no problems for polyp detection.

Overall, this dataset suffers from the exact problems this publication aims to tackle: only 0.85% of all images contain polyps. It is a highly imbalanced problem with an objectively low amount of samples compared to traditional deep learning settings.

Table 5.1: Morphology - Polyp's size in the Polyp WCE dataset, as reported in Laiz et al. (2020).

		Morphology			Total
		Sessile	Pedunculated	Undefined	
Size	Small (2–6 mm)	65	4	19	88
	Medium (7–11 mm)	29	4	20	53
	Large (12+ mm)	8	3	13	24
Total		102	11	52	165

CAD-CAP WCE

This public dataset was compiled during the Gastrointestinal Image ANAlysis (GIANA) challenge (Dray et al., 2018). It consists of three balanced classes: normal, inflammatory, and vascular lesion, each with approximately 600 images for a total of 1,800 images.

Although the classes are balanced, the total amount of samples is much smaller than the other supervised dataset. Thus, this set can be used to test if the SSL process has captured enough rich information to avoid overfitting.

5.5.2 Implementation Details

We performed all the experiments on one NVIDIA Titan Xp GPU, implementing the entire architecture in TensorFlow 2.4. The backbone network, a ResNet-50, was initialized using the Imagenet trained model, while the projection layers were randomly initialized.

Preprocessing

All data, including the used in pretrain and finetune, was processed using standard data augmentation techniques during the training phase, such as color jittering, grayscale conversion, and random rotations and flips.

Only RGB channels are used during all stages, keeping the images' size at 256 by 256 pixels downsizing them using bilinear interpolation without antialias when needed. We also introduced a mask with a radius of 128 pixels to eliminate any artifacts present at the borders of the images, making sure that no specific noise or patterns could identify either a dataset or a particular video.

For our finetune step, as is customary in the field due to the low number of images, the use of data augmentation is mandatory to avoid overfitting. We found that not introducing this same augmentation on the pretrain step negatively affected our final classification results. Thus, all sections below assume the use of data augmentation techniques for training. During evaluation no prepocessing, other than resizing, is done to the data.

Self-supervised learning

The unlabeled Generic WCE videos were used as training data during this stage. The network was optimized using stochastic gradient descent, without momentum, for a total of 21,000 batches with 72 images each (about 2 hours and 30 minutes on our GPU). In our best-performing configuration, the network processes 21,000 sequences. The learning rate was fixed to 0.1, and was divided by 5 every 4,300 iterations. Throughout the process, we used an L2 weight decay of 0.0001. We experimented with multiple values, reaching the same conclusion as SimCLR (Chen et al., 2020a), whereas any low value helps regularize the embedding pre-projection. Finally, we used a batch all strategy for triplet loss, with unnormalized embeddings and a margin of 0.2.

While the SSL network will be used as is, after training with the Generic WCE videos, it is required to find the best set of hyperparameters. To such means, a procedure has been devised. For a particular set of hyperparameters, the network is normally trained, then finetuned over the polyp dataset, and finally evaluated using Area Under the ROC Curve (AUC) computed from Receiver Operating Characteristic Curve (ROC). Here, the polyp dataset is only used as a proxy to evaluate how the hyper-parameters perform, and not as a proper evaluation of the downstream task. For instance, this procedure uses a five-fold cross-validation over randomly selected samples from the Polyp WCE videos, whereas the downstream task will be evaluated with complete videos.

Supervised learning

For each of the two supervised datasets, Polyp and CAD-CAP, the entire pretrained network was finetuned with a linear classifier on top of the learned representation. All datasets were equally trained with a learning rate of 0.01, decaying it by 10 every 1,500 iterations for a total of 4,500 steps.

5.5.3 SSL Hyperparameters

We first performed experiments to choose the sequences' length N, window size w, and whether multiple videos should be used in a single batch or not. Due to our available GPU memory, we could fit at most 72 images in a single batch, which set an upper bound to N. We designed several models, see Table 5.2, to select the best performing combination. Although the results show no statistically significant difference among some, it can be observed that sequences of 72 images, where all images come from the same video, tend to give better results. Sampling from one video or multiple at once, within a set sequence and window size, has a lower effect on the results than the length of the sequence. Due to hardware limitations, further combinations could not be tested. For instance, it is encouraged to try whether multiple sequences of 72 images are beneficial for a particular downstream task.

Most images will be relatively similar and close when using a continuous stream of 72 images. Therefore, triplets formed for TL will consist of hard negatives, namely from samples that are difficult to distinguish. Oppositely, mixing several short sequences in a single batch will produce negatives that are too easy to distinguish from their anchors.

Table 5.2: Hyperparameters tested during the self-supervised training, combining different Sequence Sizes (N) and Window Sizes (w). Resampling indicates that, in a single batch, all sequences come from the same video. Note that resampling only makes sense if N is smaller and multiple of the batch size.

Sequence Size	Sequences per Batch	Window Size	Resample	AUC (%)
9	8	3	No	93.51 ± 1.35
9	8	3	Yes	93.23 ± 1.78
9	8	6	No	93.49 ± 1.31
9	8	6	Yes	93.81 ± 2.12
18	4	3	No	93.68 ± 1.97
18	4	6	No	93.47 ± 1.11
18	4	6	Yes	92.91 ± 2.70
18	4	9	No	93.42 ± 1.62
18	4	9	Yes	93.62 ± 1.63
72	1	6	–	94.12 ± 1.35
72	**1**	**9**	–	$\mathbf{94.60 \pm 1.15}$
72	1	18	–	94.14 ± 2.12
72	1	32	–	94.53 ± 0.96

We believe this added difficulty, albeit making the training process slower, helps the network extract more meaningful information of the images. Thus, richer embeddings are produced, which can then perform better in later downstream tasks. For future experiments, N was fixed to 72, obtained continuously from a single video, and w to 9 images.

Following, we pinpointed the benefits of adding projection layers. We verified, as can be observed in Table 5.3, whether adding these additional parameters during the pretraining phase yielded better results during polyp detection. It is of particular importance to remark that any projection layer added is then removed during the second phase, thus the same number of parameters is kept regardless of the choices made here.

Table 5.3: Study of the effect of adding several projection layers with a varying number of parameters. Each projection layer consists of a ReLU activation followed by a dense layer. All dense layers have the same amount of parameters (dimensionality).

Projection Layers	Projection Dimensionality	AUC (%)
0	–	92.97 ± 1.19
1	128	93.02 ± 1.39
2	128	94.09 ± 1.28
3	**128**	$\mathbf{94.60 \pm 1.15}$
3	256	93.56 ± 1.53
6	128	93.85 ± 1.80

Particularly, the optimal combination for our particular task seemed to be at 3 layers,

each of 128 parameters, which yields a substantial improvement compared to using none and outperforms more complex solutions.

After finding the set of hyper-parameters that performs best, all models used for hyper-parameter evaluation are discarded. Downstream tasks are finetuned with the SSL network trained with the Generic WCE videos dataset, with $N = 72, w = 9$, and 3 projection layers with 128 parameters each.

5.5.4 Results

In this subsection, first the quality of the embeddings learned during the self-supervised learning is evaluated. Then, we explore the results obtained with two downstream specific tasks.

SSL embeddings

As stated, our SSL process aims to learn rich embeddings. To such end, we use the temporal sequences extracted from WCE videos to make the network learn when two images are close or not in the video. It is expected that two embeddings of consecutive images are similar.

Taking into account we measure similarity with Euclidean distance in the TL function, two embeddings are considered close if their distance is relatively near the margin parameter, or distant otherwise. As can be seen from Figure 5.3, the network successfully distinguishes not only images that are completely different but also correctly represents images that are similar while not being consecutive.

Similarly, some samples are close to frames of other videos while maintaining evident similarities, which serves to justify that the network has not learned features specific to a video, but, rather, it has trained for rich information. Our time-based contrastive learning implicitly enables the model to identify similarities between different videos with similar events, which is vital for SSL, as the finetune process needs this augmented information to properly function.

To further validate the embeddings, we obtained a t-SNE representation (Van Der Maaten and Hinton, 2008) of one WCE video. As can be seen in Figure 5.4a, frames that are visually close, containing similar structures and colors, are densely packed in the same area of the representation. This indicates that their embeddings are also close, verifying that the network has learned our contrastive metric successfully. Likewise, the network has learned that images that are close in the video, are naturally similar, Figure 5.4b. The smooth gradient of colors, following the *viridis* scheme, along with the clusters of similar colors further indicate that similar images have similar embeddings.

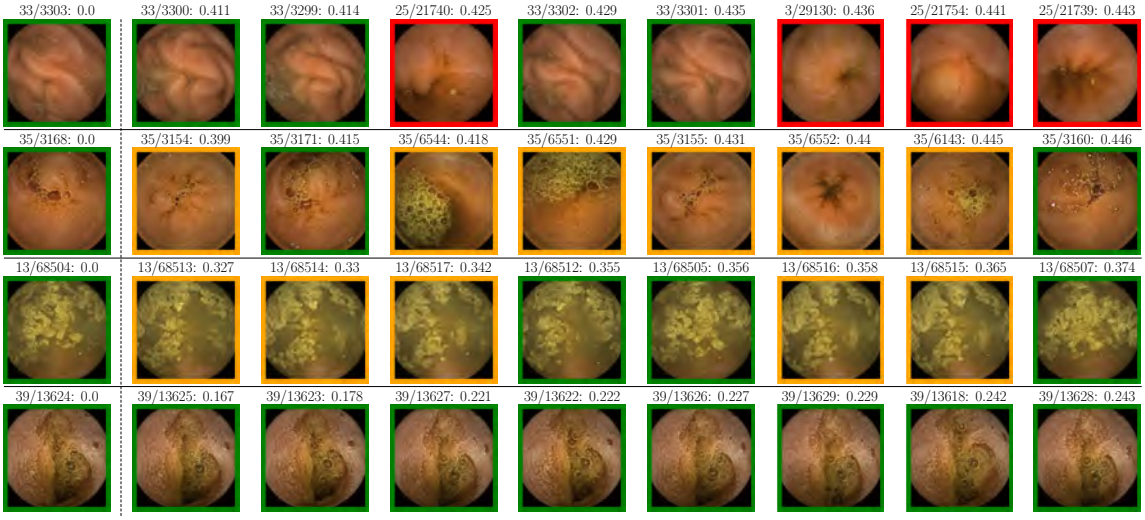
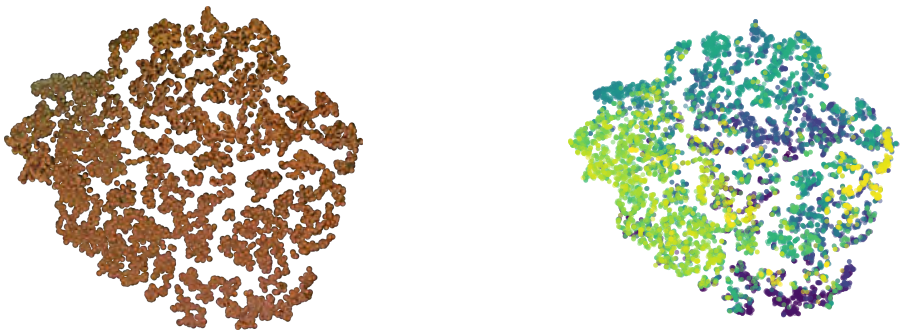


Figure 5.3: Given samples from the test set, shown in the first column, each row represents other samples in the set sampled by distance in the embedding space. Each image is titled as *video/frame: distance*, and framed in red if they come from a different video, orange if it is the same video, and green if, additionally to being in the same video, they are within w distance.



(a) Each embedding is represented with its corresponding image.

(b) Each embedding is colored according to its position in the video, following the *viridis* scheme. Images at the start of the video appear yellow, gradually turning purple as they get to the end.

Figure 5.4: t-SNE of the embeddings post-projections obtained from one WCE video after the pretrain phase. The representation shows (a) that visually alike images have close embeddings, and (b) that order is preserved.

Polyp dataset

Following previous work from Laiz et al. (2020), we abandon traditional metrics used in polyp detection. Accuracy, for instance, is a skewed metric under such data imbalances, favoring the class with most examples in detriment to the overall performance. Thus,

as proposed in their publication, we adopt AUC ROC as the primary metric. Moreover, following the same procedure in Laiz et al. (2020), sensitivity at set specificity thresholds, namely 95%, 90%, and 80%, are also reported. Not only are they robust towards imbalance, but most importantly they provide helpful information regarding the number of images a physician needs to check to obtain a certain level of performance in polyp detection. For instance, this metric gives a measure of how many polyps would be detected if a percentage of negatives was discarded based on the classifier.

To ensure that similar images, which are commonly found in sequential frames in videos, are not present in both train and evaluation simultaneously, we split the dataset based on whole videos. Consequently, a patient can only be found either in train or evaluation, but never in both. Failing to do so would overestimate the performance, producing better results while probably failing to generalize with new data.

The baseline for this particular task, further referred to as Imagenet, uses a ResNet-50 preinitialized with Imagenet and trained on this same dataset. Unlike our model, the Imagenet model uses no SSL nor any contrastive loss. A more advanced model, TL_{BA} as trained in Laiz et al. (2020), introduces a TL to the previous model. Finally, the state-of-the-art contrastive learning architecture SimCLR (Chen et al., 2020a), is also compared.

Every result, as seen in Table 5.4, is reported as the mean value and standard deviation obtained from a 5-fold cross-validation. Each evaluation set is done with whole videos, not individual samples. Also, each fold is finetuned and evaluated independently, starting from exactly the same initial values taken from our pretrained network.

Table 5.4: Performance comparison of several methods with the same parameter count. Imagenet refers to a ResNet-50 pretrained on the imagenet dataset and then finetuned with a cross-entropy loss over our dataset. SimCLR has been trained with NT-Xent as per Chen et al. (2020a). $\mathrm{TL_{BA}}$ is equivalent to Imagenet but trained with an additional triplet loss. Ours is the self-supervised network.

	AUC	**Sensitivity %**		
Model	**(%)**	**Spec. at 95%**	**Spec. at 90%**	**Spec. at 80%**
Imagenet	82.85 ± 5.72	37.75 ± 9.12	51.49 ± 11.09	66.71 ± 12.15
SimCLR (Chen et al., 2020a)	92.76 ± 1.62	68.13 ± 6.37	76.92 ± 5.40	87.91 ± 3.94
$\mathrm{TL_{BA}}$ (Laiz et al., 2020)	92.94 ± 1.87	76.68 ± 4.93	82.86 ± 4.78	88.53 ± 3.76
Ours	$\mathbf{95.00 \pm 2.09}$	$\mathbf{80.16 \pm 6.97}$	$\mathbf{86.31 \pm 6.20}$	$\mathbf{92.09 \pm 4.63}$

Adding any kind of contrastive losses, as can be seen from TL_{BA} and SimCLR in Table 5.4, already provides a significant boost over the baseline of 9.91% and 10.02% on the AUC score, respectively. Furthermore, our method based on SSL outperforms the former models by 2.24% and 2.06%, respectively, reaching an AUC score of 95.00%. A detailed view of the ROC curve is provided in Figure 5.5, where our model can be seen outperforming the rest, achieving higher true positive detections with a lower false positive rate. This significant improvement can be observed across all metrics, meaning SSL and our particular time-based

contrastive learning can extract information that remains otherwise hidden or ignored. Of particular interest are the improvements in the sensitivity at different specificity levels, as shown in Table 5.4. Our method can give a notable increase in the number of polyps correctly classified when discarding varying amounts of negatives.

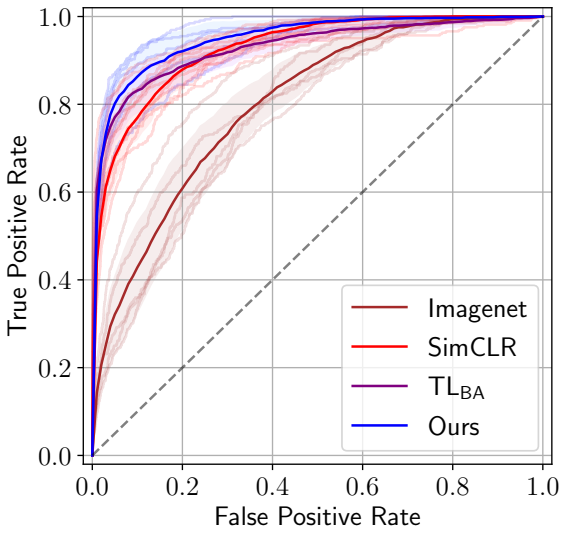
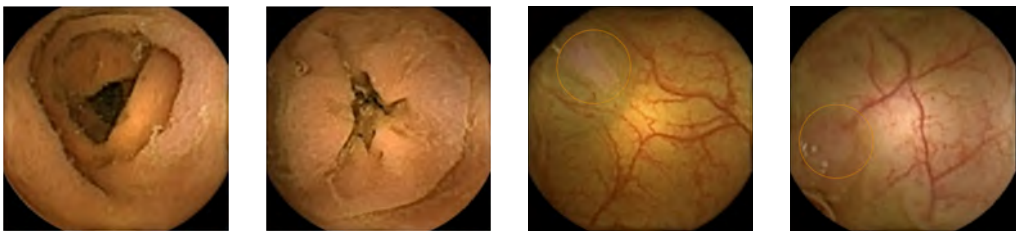


Figure 5.5: Receiver Operating Characteristic Curve (ROC) curve for the four models tested for the polyp dataset. Each cross-validation split is shown in lighter versions its corresponding model color, the mean ROC value is outlined in darker color, and the standard deviation is provided as the background shade. True Positive Rate indicates the percentage of polyps correctly identified, while False Positive Rate is the percentage of non-polyps misclassified as polyps

Another approach to validation, aside from the quantitative analysis above, is to inspect and visualize the results. In other words, performing a qualitative validation of the results by examining where the model is performing correctly and where it is failing. Miss-classified non-polyp images would add more work to the physician due to having to unnecessarily check false positives. However, not showing a polyp frame because the system has falsely classified it as negative can have a devastating effect, with implications much severe than its counterpart case. Figure 5.6 depicts two examples of the mentioned cases. It can be seen that the network fails in especially tough cases, where the polyp would be hard to be seen even for a physician. The polyps have been circled for the reader to identify where they are. False positives occur in zones with a more pinkish tone, characteristic of polyps, and always in rugged and wrinkled surfaces, which could explain why the network is mistaking them for polyps.



(a) (b) (c) (d)

Figure 5.6: Random samples from the test set. a) and b) shows two false positives, images inaccurately classified as polyps. c) and d) depicts two false negatives. The polyps have been circled to help with their identification.

CAD-CAP dataset

Following the procedure established in Guo and Yuan (2020), we have split the data into 4 sets and performed a 4-fold cross-validation. As per the original challenge (Dray et al., 2018), we report in Table 5.5 the per-class Matthews Correlation Coefficient (MCC) and F1 scores, and the overall accuracy as p_0.

A naive implementation, using a ResNet-50 and without SSL, fails to correctly classify a significant portion of the data, achieving only a 69.98% accuracy. However, adding SSL to this same model and using the method we propose in this publication, immediately boosts every metric by more than 20%. Our implementation reaches a total of 92.77% accuracy without any change to the architecture.

Further, we compare our results with those reported by Guo and Yuan (2020), the current state-of-the-art model for CAD-CAP. They handcrafted a network for this dataset and provide six baselines and one additional model that uses semi-supervision to improve the results. With respect to the baselines, our model obtains higher scores across most metrics, as can be observed in Table 5.5. We also attain comparable results to their best implementation, which has a semi-supervised phase training over 1807 unlabeled images provided by CAD-CAP that we do not use.

These results, from a clinical point of view, provide a positive step towards the simultaneous detection of several pathologies. For instance, results show that standard models that do not rely on SSL tend to accurately classify normal images, but miss a notable amount of the positive classes. Their SSL counterparts, however, keep the approximate same level of detection for normal samples, while they significantly boost the ability to detect inflammatory and vascular lesions. This encouraging accuracy would enable bringing physicians and experts into the loop, further developing the model and producing a CADx system capable of aiding in diagnosis.

Table 5.5: Per class and overall results of various methods in GIANA. ResNet is the same architecture as Ours but without the SSL step. Baseline 1 and 6 refer to the baselines reported by Guo and Yuan (2020), while the model with the same name is their semi-supervised performing implementation. Here $\mathbf{p_0}$ indicates the mean accuracy across all classes.

Method	Class	F1-Score (%)	MCC (%)	$\mathbf{p_0}$ (%)
ResNet	Normal	73.28 ± 3.57	60.58 ± 5.44	69.98 ± 1.35
	Inflammatory	65.19 ± 2.95	55.86 ± 1.77	
	Vascular	70.79 ± 4.60	65.35 ± 3.80	
Baseline 1 (Guo and Yuan, 2020)	Normal	94.92 ± 0.71	92.37 ± 1.07	84.99 ± 0.80
	Inflammatory	79.24 ± 1.55	68.72 ± 2.15	
	Vascular	80.75 ± 1.65	71.49 ± 2.57	
Baseline 6 (Guo and Yuan, 2020)	Normal	96.41 ± 0.84	94.61 ± 1.26	91.92 ± 1.71
	Inflammatory	88.98 ± 2.13	83.44 ± 3.24	
	Vascular	90.27 ± 2.78	85.75 ± 3.73	
Ours	Normal	95.00 ± 1.13	92.57 ± 1.66	92.77 ± 1.20
	Inflammatory	89.87 ± 1.65	84.99 ± 2.46	
	Vascular	90.26 ± 1.76	85.78 ± 2.37	
Guo and Yuan (2020)	Normal	97.41 ± 0.45	96.10 ± 0.69	93.17 ± 1.14
	Inflammatory	90.30 ± 1.56	85.43 ± 2.24	
	Vascular	91.69 ± 1.21	87.78 ± 2.06	

5.6 Conclusion

In this work, we propose an SSL method that leverages the information in the temporal axis of WCE videos to obtain rich embeddings. Our method introduces a pseudo-labeling process that enables time-based contrastive learning, forcing frames close in a video to be represented by similar embeddings.

We demonstrate that using this process yields better results in subsequent models specializing in domain-specific tasks. Using the SSL model to classify polyps shows an increase in successful polyp detection, achieving a 95.00% AUC, a significant improvement over existing methods. Similarly, we test the method to detect several events in the GIANA dataset, obtaining comparable results to state-of-the-art models while offering reduced complexity and a more general approach.

It is a limitation of our SSL method that the data used during the pretrain stage must come in a video format. This makes it directly applicable for WCE datasets, but would require adaptation for other medical fields. The pretaining phase is also limited by the hardware capacity, specially so since results show that longer sequences produce richer embeddings. If deployed as a CADx system, our work would only require individual samples and appropriate hardware to run.

Thus, we claim that using SSL when leveraging temporal information is beneficial for WCE models. Most importantly, the method imposes no requirements for the dataset

used during the supervised phase, effectively tackling the classical problems commonly encountered in medical imaging: low amounts of data—specially labeled—and severe class imbalances.

Overall, we strongly believe the method is a good step towards better models that empower CADx models in medical interventions. For instance, a higher rate of polyp detection would decrease the time spent by physicians revising WCE videos, allowing more accurate diagnosis in shorter amounts of time.

Future work could focus on exploring other SSL architectures that might boost the downstream tasks' performance, while exploring other hyper-parameters settings and sampling mechanisms. Moreover, expanding the method to other WCE domains and other medical fields would also be of high interest.

Acknowledgments

This work was partially founded by MINECO Grant RTI2018-095232-B-C21, SGR 1742, Innovate UK project 104633, and by an FPU grant (Formacion de Profesorado Universitario) from the Spanish Ministry of Universities to Guillem Pascual (FPU16/06843). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp Pascal GPU used for this research.

Chapter 6

Paper III

Anatomical landmarks localization for capsule endoscopy studies

Pablo Laiz[a], Jordi Vitrià[a], Pere Gilabert[a], Hagen Wenzek[b], Carolina Malagelada[c], Angus J. M. Watson[d] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

c) Digestive System Research Unit, University Hospital Vall d'Hebron, Barcelona, Spain

d) Department of Surgery, Raigmore Hospital, Inverness, UK

In Computerized Medical Imaging and Graphics, Elsevier, Volume 108:102243, 2023. ISSN 0895-6111

doi: 10.1016/j.compmedimag.2023.102243

Impact Factor Journal: 7.422
Q1 Biomedical Engineering (15/98)

Contents

Motivation and Context

The development of CAD systems has led to significant advancements in medical imaging. In particular, their application for pathology identification in WCE studies has gained much attention, due to their potential to reduce the time required to complete the video analysis. While the identification of anatomical landmarks has not received the same amount of attention as other CAD applications, it is another approach to accelerate the reading process. Moreover, it is also used to limit the region of the GI tract where lesions have been found, and it may be considered a preliminary step for algorithms targeting one specific organ (Koulaouzidis et al., 2021).

As discussed in Section 2.2.2, identifying the entrance and exit of an organ in the digestive tract using WCE is a complex task. The reason behind that is the presence of numerous similar-looking structures and the capsule movement. To address this issue, the following study proposes a method to automatically localize relevant anatomical landmarks that can be used in clinical practice. The system not only classifies the images from WCE videos, as being inside or outside the studied organs but, also identifies the entrance and exit frames, which correspond to the anatomical landmarks.

To overcome the aforementioned challenges, the proposed method employs a three-step approach. Firstly, a CNN is used to extract a low-dimensional representation of the images. The obtained embeddings and timestamps of each image are used to predict the probability of an image belonging to an organ. Second, a signal is generated for every video by leveraging the probabilities derived from each image. Since the signal can be noisy, the probabilities are smoothed using context data and bidirectional RNNs with motion and time information. Finally, the boundaries of the organs are identified by solving a minimization problem.

The proposed method is evaluated using three different datasets of the small bowel and the large intestine. The results showed that it outperformed the baseline and state-of-the-art methods in WCE videos in both tasks: frame classification and anatomical landmark identification. Hence, the proposed approach has the potential to reduce the time required for screening and improve the accuracy of pathological localization in clinical practice.

Anatomical landmarks localization for capsule endoscopy studies

Pablo Laiz[a], Jordi Vitrià[a], Pere Gilabert[a], Hagen Wenzek[b], Carolina Malagelada[c], Angus J. M. Watson[d] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

b) CorporateHealth International ApS, Denmark

c) Digestive System Research Unit, University Hospital Vall d'Hebron, Barcelona, Spain

d) Department of Surgery, Raigmore Hospital, Inverness, UK

6.1 Abstract

Wireless Capsule Endoscopy is a medical procedure that uses a small, wireless camera to capture images of the inside of the digestive tract. The identification of the entrance and exit of the small bowel and of the large intestine is one of the first tasks that need to be accomplished to read a video. This paper addresses the design of a clinical decision support tool to detect these anatomical landmarks. We have developed a system based on deep learning that combines images, timestamps, and motion data to achieve state-of-the-art results. Our method does not only classify the images as being inside or outside the studied organs, but it is also able to identify the entrance and exit frames. The experiments performed with three different datasets (one public and two private) show that our system is able to approximate the landmarks while achieving high accuracy on the classification problem (inside/outside of the organ). When comparing the entrance and exit of the studied organs, the distance between predicted and real landmarks is reduced from 1.5 to 10 times with respect to previous state-of-the-art methods.

Key words: Anatomical Landmarks; Deep learning; Wireless Capsule endoscopy; Organ detection.

The paper has been re-typeset to match the thesis style.

6.2 Introduction

Wireless Capsule Endoscopy (WCE) (Iddan et al., 2000) is a medical procedure designed to visualize the entire digestive tract through a swallowed vitamin-size capsule, which is propelled by peristalsis via the esophagus, stomach, small intestine, and large intestine (also referred to as colon). WCE offers several benefits to patients, clinicians, and the healthcare system in comparison with traditional endoscopic procedures. It does not require sedation, is less likely to cause discomfort, and presents fewer potential complications. It also minimizes the needed medical resources compared to the standard screening technique (Darrow, 2014).

Currently, in several countries, small bowel WCE is used as the first indication for obscure Gastrointestinal (GI) bleeding, Crohn's disease, and to a lesser extent, screening in polyposis syndromes, celiac disease, or other small bowel pathologies (Trasolini and Byrne, 2021). Meanwhile, colon WCE is increasingly recognized as a reliable option for polyp detection, investigation of inflammatory bowel diseases or completion of an incomplete colonoscopy (Yung et al., 2016; Koulaouzidis et al., 2021).

Unfortunately, the adoption of this technique is below the initial expectation, mainly because WCE: 1) does not admit any surgical intervention; 2) does not provide the exact location of the pathology or organs; and 3) generates recordings with thousands of frames that must be reviewed by experts, entailing a complex and time-consuming task. Even an experienced reader may require at least an hour to analyze the data of a single patient (Maieron et al., 2004; Rondonotti et al., 2020; Dokoutsidou et al., 2011).

Artificial Intelligence (AI) methods are being employed in several solutions to overcome WCE limitations and accelerate the reviewing process for readers. While most studies have been centered on detecting images with abnormalities, such as polyps, tumors, bleeding, or ulcers, few of them are focused on localizing the findings or the anatomical landmarks.

In the clinical field, the localization of anatomical landmarks and abnormalities represents a problem of particular interest as it is essential to guide gastroenterologists during the screening and to take clinical decisions (Iakovidis and Koulaouzidis, 2015). Indeed, the localization of these landmarks is one of the first tasks carried out by the readers and is required to perform a complete exploration (Koulaouzidis et al., 2021).

In this paper, we propose a deep learning method for automatically localizing relevant anatomical landmarks to be used in the clinical routine with different capsule endoscopy devices. The aim of this work is to reduce the average time required to complete the clinical routine which typically takes approximately 25 minutes by a specialist reader (Iakovidis and Koulaouzidis, 2015). To reach this purpose, the method focuses on detecting the end of the pylorus and the ileocecal valve, which delimit the small bowel. For the large intestine, the points of interest are the first cecal and last rectal images. Moreover, the last rectal image ensures the proper identification of the farthest point of capsule progression. These

landmarks are illustrated in Figure 6.1 bordered by a green dashed line. Each one of them can be visualized in multiple depending on the orientation of the capsule. Though, in some cases, they can be hidden by GI content, which increases the complexity of the task.

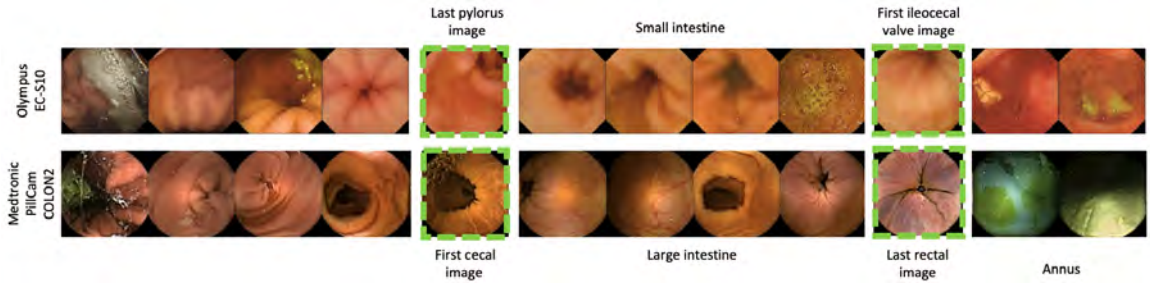


Figure 6.1: Illustration of random frames from two GI tracts. The first sample is recorded with the Olympus EC-S10, whereas the second one is obtained with Medtronic PillCam COLON2. The corresponding landmarks of the small bowel (first sample) and the large intestine (second sample) are bordered by a green dashed line.

First, our system aims to identify all the images between the landmarks using video frames, and additionally, timestamps and motion information that have not been employed in previous studies. Subsequently, the model recognizes the first and last image belonging to the studied organ. The obtained results show that, by providing extra knowledge to the network, the performance of the system increases compared to the state-of-the-art methods and may reduce the average time to complete the clinical routine.

The paper is organized as follows: initially, an overview of the related work in the field is given. Then, our method is presented in detail explaining the key steps, followed by the experimental setup, where the three used databases and metrics are introduced. After that, the results of the experiments are extensively exposed in a quantitative and qualitative manner to prove the performance of the method. Finally, the conclusions and future work are discussed.

6.3 Related Work

The related work can be divided into two main categories: *ad hoc* techniques and deep learning models. First, the traditional statistical methods and machine learning techniques are reviewed. Then, deep learning methods focused on organ classification are summarized.

Berens et al. (2005) were the first to propose a solution for the detection of anatomical landmarks. They employed hue saturation chromaticity histograms to distinguish the stomach, intestine, and colon tissues. Lee et al. (2007) made use of intestinal contractions to locate the boundaries of the organs or unusual events such as intestinal juices, bleeding, and rare capsule movements. Mackiewicz et al. (2008) described the use of color image analysis to discriminate between the esophagus, stomach, small intestine, and colon. Haji-

Maghsoudi et al. (2012) proposed an algorithm to classify the same organs using static and non-static features. Li et al. (2015) reported a method that draws a dissimilarity curve implementing the color feature to locate the boundaries between the stomach, small intestine, and large intestine. In these methods, the system performance is assessed as the frame distance (error) between the point in the video where the boundary was manually annotated by a clinician and the one selected by the algorithm.

The latest methods to discriminate between organs are based on deep learning techniques. Zou et al. (2015) proposed a network called DCNN-WCE-CS to classify the digestive organs from WCE images by recognizing high-level semantic features. The network was built with three convolutional layers and a dense layer to classify. Chen et al. (2017a) presented two different systems, O-CNN and TO-CNN. The former consisted of a standard Convolutional Neural Network (CNN), "AlexNet", whereas the latter additionally integrated temporal information employing Hidden Markov models. Adewole et al. (2020) compared four state-of-the-art Deep Neural Networks (DNNs) to detect the anatomical parts within the GI tract. Zhao et al. (2021) designed a three stages method to detect the boundaries of the small bowel. The method explores long-range temporal dependency with a transformer module, which captures the temporal inter-frame dependencies in short sequences. To locate the starting and ending of the organ, a search algorithm is applied. Finally, Son et al. (2022) proposed a system based on a DNN with temporal filtering (a combination of median and Savitzky-Golay filters) on the predicted probabilities. To detect the boundaries, the method considers the minimum and maximum frame index predicted as small bowel. Although in terms of classification, deep learning methods outperform the obtained results with the extraction of handcrafted features, only Zhao et al. (2021) and Son et al. (2022) apply thresholding techniques to identify the boundaries of the small bowel. To the best of our knowledge, all the studies were performed using private datasets and with only one type of capsule.

6.4 Method

Our method aims to localize the anatomical landmarks from WCE videos. An overview of the employed strategy is illustrated in Figure 6.2. To achieve the primary purpose, the main steps are: 1) Develop a deep learning model to predict the probability of each image to belong to the area of interest, the small bowel or the large intestine; 2) Smooth and mitigate any noisy behavior of the probabilities with extra information (temporal and motion data); 3) Predict the boundaries using a rectangular pulse function by a minimization problem.

6.4.1 Step 1: Probability prediction

Let $x_i \in X$ be an image, where x_i is the i^{th}-frame of a WCE video X, and $f(\cdot)$, a DNN architecture. The low representation of the image x_i is defined as $x_i' = f(x_i) \in \mathbb{R}^{2048}$. The

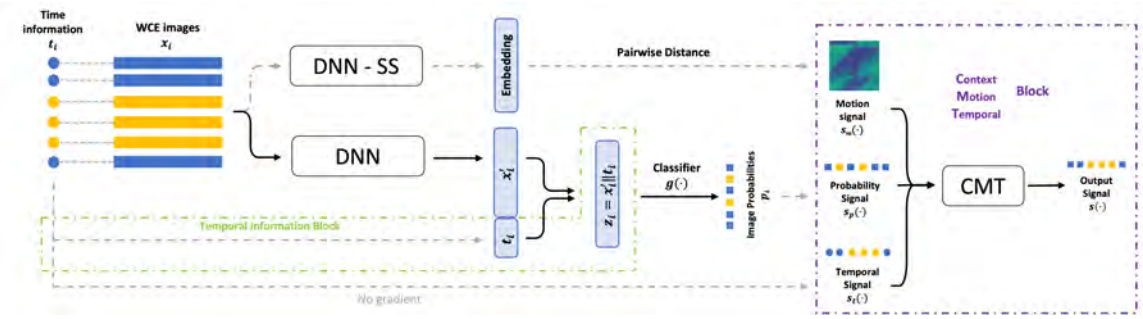


Figure 6.2: Overview of the proposed system. The input of the network consists of a sequence of images and their temporal information. The main architecture is a DNN concatenated with the temporal and CMT blocks. The output of the model is a smooth signal with low noise.

vector x'_i is extended by adding a new feature containing the temporal information of the frame, $z_i = x'_i \parallel t_i \in \mathbb{R}^{2049}$, $t_i \in [0, 1]$.

The added time-related feature, t_i, is based on the image timestamp and exists for each i. Each image is mapped to a value between zero and one according to:

$$t_i = \frac{\text{timestamp}_i}{\text{video length}} \tag{6.1}$$

where $timestamp_i$ represents the time (in seconds) of the i^{th}-frame in the video. This equation normalizes all the video lengths and provides the temporal position with respect to the entire video.

The WCE advances through the GI tract recording all organs in a continuous manner. It is worth remarking that although the camera might go back and forth, it remains in the same organ. This allows the model to create a relationship between time and organs. The temporal feature added by our system allows the model to discard erroneous predictions in different sections of the video.

Then, using a linear classifier $g(\cdot)$ and the extended vector z_i, the probability p_i of each frame is inferred:

$$p_i = g(z_i) = g(x'_i \parallel t_i) = g(f(x_i) \parallel t_i) \tag{6.2}$$

6.4.2 Step 2: Smoothing the probabilities

Figure 6.1 contains some examples of frames where the mucosa of the digestive tract is hidden by noisy content (Chen et al., 2017a) like bile, bubbles, residues, and liquids. In those frames, the network may yield senseless probabilities. To mitigate this undesirable behavior, it is important not only to properly analyze a still image but the entire sequence. Furthermore, if the context analysis is complemented with the movements of the capsule within the intestine and the temporal information, the developed model can further decrease this erratic behavior.

Given all the frames from a video, the capsule movement signal $s_m(\cdot)$ is obtained by estimating the distance between frames. To calculate them, the time-based self-supervised network, $f_{ss}(\cdot)$, proposed by Pascual et al. (2022a) is employed to obtain the embedding for each image x_i, $e_i = f_{ss}(x_i)$. The network generates similar representations for images that are close in time, i.e., consecutive frames from the same sequence have similar embeddings. For images from different sequences, the network yields distant image representations. Then, the Euclidean pairwise distance is computed between the embeddings to obtain the matrix M. The values of M are an approximation of the motion between two frames. Small values of the matrix are caused by small movements of the capsule, while high values mean the opposite.

The visualization of this matrix shows contraction patterns of the GI tract and can suggest where the camera might be located. Because of the length of the video, the complete matrix is difficult to visualize. Therefore, it is simplified as a figure containing the i^{th}-frame, centered in the middle of each row and their 500 nearest temporal neighbors, all of them represented as pixels. The color of each one is the distance between the frame and the central one. The darkest points correspond to small distances, implying that the capsule hardly moves. While, the lighter pixels point out larger distances, which entails a drastic movement of the capsule. Figure 6.3 contains three samples of the capsule movement codified as an image. Each one is shown in four parts: the beginning of the video, the first landmark, a random segment of the organ, and the second landmark. The frames containing the landmarks annotated by the experts are represented with black dashed lines.

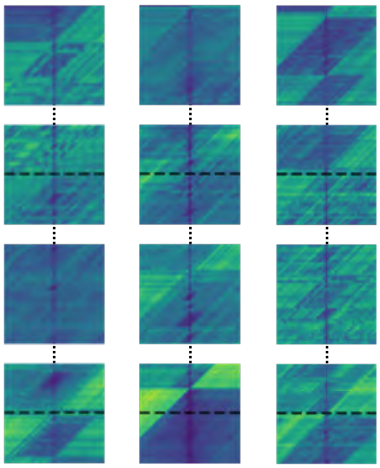


Figure 6.3: Movement visualization of the capsule in different dataset videos: a) *Kvasir-Capsule* dataset, b) *VH* dataset and, c) *Capri* dataset. Each row corresponds to the beginning of the video, the first landmark, a random part of the organ, and the second landmark. Inside each patch, the x-axis represents the relationship of the central frame to the other frames, and the y-axis contains the frames in chronological order.

The sequential analysis is performed in the context - motion - temporal (CMT) block, which smooths the probabilities of those frames with senseless values by combining neighborhood probabilities, motion, and time information. The use of the CMT block is a paradigm shift which works with probabilities and information from the whole video encoded in three signals:

- Probability signal, $s_p(\cdot)$, is obtained by concatenating the probabilities inferred in each frame of a video:

$$s_p(i) = g(f(x_i) \parallel t_i), \quad \forall i \mid x_i \in X \tag{6.3}$$

- Motion signal, $s_m(\cdot)$, is obtained by using the normalized i^{th}-row of the matrix M.

- Temporal signal, $s_t(\cdot)$, is obtained by concatenating the time information of each frame of a video:

$$s_t(i) = t_i, \quad \forall t_i \tag{6.4}$$

These signals are concatenated vertically to generate a matrix of size $3 \times \textit{video length}$. To calculate the output signal, $s(\cdot)$, a small network called $\text{CMT}_w(\cdot)$ is used. It is composed of two layers of bidirectional LSTM cells and one dense layer over w consecutive frames. This is formalized as:

$$s = \text{CMT}_w(s_p \parallel s_m \parallel s_t) \tag{6.5}$$

The window size hyper-parameter, w, is a natural odd number that must be determined to achieve optimal results. The overview of this block can be seen in Figure 6.4.

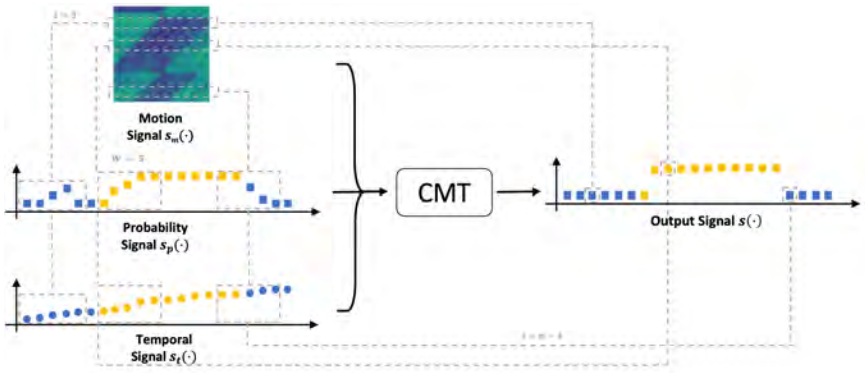


Figure 6.4: Overview of the proposed CMT block with $w = 5$. The input of the block is the different signals extracted from processing a WCE video: the probability signal s_p, the motion signal m and the temporal signal s_t. The output signal s is obtained after combining the given information.

6.4.3 Step 3: Boundaries prediction

Finally, a simple but efficient technique is employed to identify the landmarks of the WCE video using the inferred probabilities of each image belonging to an organ. A minimization problem, $\varphi(\cdot)$, is solved over the output signal to identify the boundaries of the organ, as it is shown in Figure 6.5. Let $V(t)$ be the rectangular pulse function:

$$V(t) = u(t-a) - u(t-b) \tag{6.6}$$

where $u(t)$ is the unit step function defined as:

$$u(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases} \tag{6.7}$$

and a and b are the limits where the function $V(t)$ has value one. To identify the first and last frame of the organ, the distance between the output signal $s(i)$ and the rectangular pulse $V(t)$ is minimized by finding the best a and b values:

$$\begin{aligned} &\underset{a,b}{\text{minimize}} \sum_{i=0}^{a-1} s(i) + \sum_{i=a}^{b-1} 1 - s(i) + \sum_{i=b}^{n} s(i) \\ &\text{s.t.} \quad a < b \end{aligned} \tag{6.8}$$

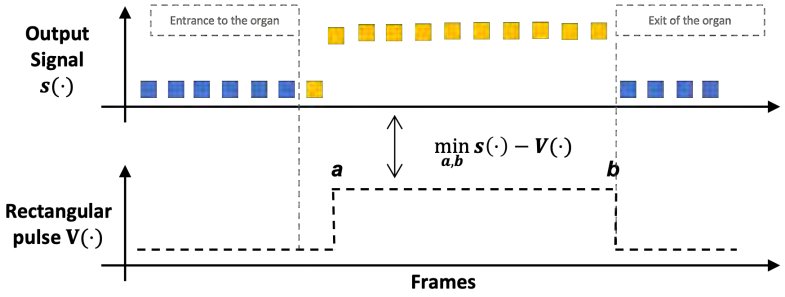


Figure 6.5: Overview of the minimization problem, given the output signal of the video and the rectangular pulse function required to solve Equation 6.8. Grey lines correspond to the anatomical landmark annotated by the expert.

The optimization of the network weights is carried out using the binary cross-entropy loss to minimize Equations 6.2 and 6.5. In both cases, the real binary labels inside/outside the organ have been used to compute the cross-entropy during training.

6.5 Experimental Setup

6.5.1 Datasets

The proposed system is evaluated with one public dataset (*Kvasir-Capsule*) and two private ones (*VH* and *Capri*).

Kvasir-Capsule dataset

This public dataset was collected from 117 examinations at a Norwegian Hospital employing the Olympus Endocapsule 10 System (EC-S10) (Smedsrud et al., 2021). In our case, we only used the set of 24 videos that contains anatomical landmarks of the small bowel, specifically the pylorus and the ileocecal valve. The number of frames per video is $44K$ on average. Small bowel images represent 75.14% of the dataset. However, this dataset does not contain the temporal information of the videos.

VH dataset

The second dataset was obtained from 48 healthy volunteers. Physicians from Vall d'Hebron Hospital in Barcelona recorded all the videos using Medtronic PillCam SB3 and labeled the limits of the small bowel. The average number of frames per video is $35K$ with a mean video duration of 04:36:06. The frames between the pylorus and the ileocecal valve represent the 70.68% of this dataset.

Capri dataset

The last used database is composed of 68 colon studies from different patients. All these WCE videos were recorded with Medtronic PillCam COLON2 on behalf of the NHS Highland Raigmore Hospital in Inverness. Images from both cameras, frontal and rear, from the PillCam COLON2 are used in the experiments. The mean duration of the videos is 08:19:51 with an average of $14K$ frames. The colon images represent the 74.63% of the dataset.

6.5.2 Evaluation criteria

Models are evaluated with a two-fold stratified cross-validation strategy, following the instructions established in Smedsrud et al. (2021). It is worth remarking that the stratified partitions are not based on individual frames but on individual patients. Hence, images from the same patients do not belong to different sets. Table 6.1 contains the details about each fold for each one of the used datasets.

As in previous (Zou et al., 2015; Chen et al., 2017a; Adewole et al., 2020; Zhao et al.,

Table 6.1: Overview of the records in the three datasets used in this paper. The column *#Inside* refers to those frames that are between the landmarks specified in each dataset. Respectively, column *#Outside* refers to the number of frames that do not belong to the area of interest.

Dataset	Partitions	#Patients	#Inside	#Outside	Total
Kvasir-Capsule	**Fold 0**	12	400K	160K	560K
	Fold 1	12	384K	97K	481K
	Total	**24**	**784K**	**257K**	**1M**
VH	**Fold 0**	24	602K	246K	848K
	Fold 1	24	592K	249K	841K
	Total	**48**	**1.2M**	**495K**	**1.6M**
Capri	**Fold 0**	34	347K	148K	495K
	Fold 1	34	393K	97K	490K
	Total	**68**	**740K**	**245K**	**985K**

2021; Son et al., 2022), the performance of the method in the classification task is measured with the following metrics: the Area Under the ROC Curve (AUC), Accuracy (ACC), Mean Accuracy (MACC), Specificity (SPEC), and Sensitivity (SENS).

The AUC and MACC are the most appropriate metrics for evaluating the performance of a binary classification model on imbalanced datasets. The AUC measures the model's ability to distinguish between images that belong to the target organ and those that do not, while the MACC and ACC measure the number of images that are correctly predicted. It is important to note that relying solely on SENS and SPEC rates for comparison can be problematic, as these can vary depending on the chosen cut-off thresholds.

Similar to Mackiewicz et al. (2008); Li et al. (2015); Zhao et al. (2021); Son et al. (2022), the performance of localizing the anatomical landmarks is assessed as the frame distance (error) between the image where the boundaries of the organ were manually annotated by the experts and those predicted by the system. Mean Absolute Error (MAE) and median absolute error are used to quantify the performance. Since the capsule frame rate is variable, both errors (MAE and median) are also presented as the difference in time (except in the *Kvasir-Capsule* dataset, where frame time information is not available).

It is important to note that the metrics are computed per video to avoid any bias caused by the video lengths.

6.5.3 Implementation details

TensorFlow 2.4 was used to implement the models, which were executed on a machine with an NVIDIA GeForce RTX 2080 TI and CUDA 11.0. The training process is composed of two separate stages. Firstly, the DNN with the temporal block is trained. Then, the weights

of the DNN are frozen and the CMT block is optimized.

ResNet-50 (He et al., 2016) initialized with ImageNet weights has been used as a backbone architecture for the first stage DNN. The optimization of the network was carried out with Stochastic Gradient Descent and a batch size of 256. In all the experiments, the networks were trained for $10K$ iterations. For all the datasets, the learning rate was set to 0.1 and it was decreased by a factor of 0.1 every $2K$ iterations.

All the images were resized to 128×128 pixels. In the case of the private datasets, a uniform circle mask was applied over each frame to eliminate the artifacts present at the borders of the images, ensuring that no specific noise or patterns could identify either a dataset or a particular video.

Data augmentation techniques were applied during the training phase to improve the robustness of the method. Specifically, rotations of 0, 90, 180, and 270 degrees, horizontal and vertical flips, and changes in the brightness of the images were used.

The CMT network from the second stage is composed of two bidirectional LSTM layers with 200 and 100 units, respectively. Finally, the dense layer has two neurons as output. The network was optimized with RMSprop as it is recommended in Zaman et al. (2021). The learning rate was fixed to 0.001 during $4K$ iterations. The batch size was set to 512. To find the optimal hyper-parameter window size, w, a search grid was done, and the chosen value was = 201 for the *Kvasir-Capsule* and *Capri* datasets and w = 151 for the *VH* dataset.

Since the presented datasets are statistically different, the hyper-parameter window size, w, must be chosen carefully for each case. The reported metrics in Table 6.2 are the AUC score of the model in the image classification task and the median error in the entrance, exit, and sum of both in the landmark identification task. A window size of 201 achieves the best results with the AUC metric for all the datasets. At the entrance of the organ, the smallest error is obtained with $w = 51$, whereas the lowest error in the exit is achieved with $w = 151$ in *VH* dataset and $w = 201$ in *Kvasir-Capsule* and *Capri* datasets. The same values of w are the ones that obtain the lowest error in the sum of the entrance and exit of the organs. Therefore, the hyper-parameter w chosen for *Kvasir-Capsule* and *Capri* datasets is $w = 201$ because the AUC and total median error coincide. In the case of the *VH* dataset, the chosen value is $w = 151$ since the difference between the AUC values for $w = 151$ and $w = 201$ is negligible.

6.6 Results

The results section is divided into two sets of experiments. The first one presents the performance of the classification task for each one of the datasets. The second is focused on identifying the exact frames where the capsule enters and exits the studied organ. Finally, qualitative results are shown to complement the quantitative results of the proposed system compared to methods published so far.

Table 6.2: Window size hyper-parameters tested during training. The metrics used to identify which is the best value are the AUC and the total median error obtained in a two-fold cross-validation.

Window Size	**Datasets**											
	Kvasir-Capsule				**VH**				**Capri**			
	AUC	**Median Error**			**AUC**	**Median Error**			**AUC**	**Median Error**		
		Entrance	Exit	Total		Entrance	Exit	Total		Entrance	Exit	Total
11	95.66	58.25	982.00	1040.25	98.09	46.50	266.75	313.25	99.61	4.00	1.50	5.50
51	95.47	**55.75**	693.25	749.00	98.66	**31.25**	276.75	308.00	99.79	**2.75**	1.50	4.25
75	95.53	82.75	954.25	1037.00	98.59	35.75	260.00	295.75	99.74	3.00	1.50	4.50
101	94.60	75.75	1540.50	1616.25	98.55	37.25	259.00	296.25	99.76	3.50	1.75	5.25
151	95.41	111.75	1082.75	1194.50	98.54	41.50	**210.75**	**252.25**	99.78	6.25	1.50	7.75
201	**96.00**	76.50	**487.25**	**563.75**	**98.68**	53.75	260.00	313.75	**99.79**	**2.75**	**1.00**	**3.75**
251	93.71	92.75	758.00	850.75	98.43	43.50	218.00	261.50	99.62	3.50	2.25	5.75
301	95.63	76.00	777.00	853.00	98.41	42.50	218.00	260.50	99.70	**2.75**	**1.00**	**3.75**

In all the experiments, *Kvasir-Capsule*, *VH*, and *Capri* datasets are evaluated using our proposed method, which consists of a DNN concatenated with the temporal and CMT blocks. To analyze the influence of each component, an ablation study is performed by building several additional models. Finally, our method is compared with the state-of-the-art works in each task.

For the ablation study, *ResNet* is the simplest method and is considered the baseline for the comparisons. The influence of the temporal block is evaluated with the *ResNet + Time* model, which combines the image representation obtained by *ResNet* with the timestamp of the image. The probability signals generated with the outputs of each model are used to study the contribution of the context block. Let's note that *ResNet + Time + CMT* is the proposed method. Table 6.3 contains a summary of the ablation settings of each method.

Table 6.3: Overview of the ablation settings and the name used.

Method	Ablation Settings				
	ResNet	Temp. Block	Context		
			Prob.	Motion	Time
ResNet	✓				
ResNet + C	✓		✓		
ResNet + CM	✓		✓	✓	
ResNet + CT	✓		✓		✓
ResNet + CMT	✓		✓	✓	✓
ResNet + Time	✓	✓			
ResNet + Time + C	✓	✓	✓		
ResNet + Time + CM	✓	✓	✓	✓	
ResNet + Time + CT	✓	✓	✓		✓
Proposed Method	✓	✓	✓	✓	✓

6.6.1 Image classification

The first experiment evaluates the performance of the proposed method in the image classification problem. Specifically, on *Kvasir-Capsule* and *VH* datasets, all the compared models aim to identify the small bowel frames, whereas, on the *Capri* dataset, they aim to classify the large intestine images.

Table 6.4: Comparison of the ablation study in the image classification problem for each dataset. Displayed results are the mean obtained after evaluating a two-fold cross-validation.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	ResNet + C	93.53 ± 5.51	87.24 ± 13.04	82.78 ± 10.18	73.05 ± 16.82	92.50 ± 11.90
	ResNet + CM	92.70 ± 6.22	88.47 ± 11.18	83.95 ± 9.71	74.41 ± 16.75	93.49 ± 10.66
	ResNet + CT	94.40 ± 6.22	87.65 ± 11.59	83.72 ± 11.06	75.45 ± 18.98	91.98 ± 11.32
	ResNet + CMT	95.47 ± 5.39	90.07 ± 7.25	85.18 ± 8.86	75.23 ± 17.76	95.12 ± 6.54
	ResNet + Time	92.40 ± 5.06	87.88 ± 6.04	81.16 ± 7.66	67.92 ± 15.42	94.40 ± 5.27
	ResNet + Time + C	94.91 ± 4.29	89.53 ± 6.73	**87.69 ± 7.31**	**82.11 ± 15.57**	93.27 ± 6.96
	ResNet + Time + CM	94.67 ± 5.24	89.87 ± 6.54	87.39 ± 7.48	80.83 ± 15.88	93.95 ± 6.30
	ResNet + Time + CT	**96.36 ± 3.98**	90.80 ± 5.80	87.62 ± 7.77	80.21 ± 16.73	95.03 ± 4.92
	Proposed Method	96.00 ± 4.57	**91.36 ± 5.75**	87.47 ± 7.49	78.91 ± 16.28	**96.03 ± 4.29**
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	ResNet + C	96.28 ± 6.93	93.44 ± 6.27	91.89 ± 8.01	87.31 ± 15.88	96.47 ± 3.97
	ResNet + CM	97.70 ± 4.05	93.78 ± 5.94	91.97 ± 7.86	87.15 ± 15.69	96.79 ± 3.93
	ResNet + CT	97.81 ± 3.69	92.59 ± 6.93	91.12 ± 8.37	86.10 ± 17.00	96.15 ± 4.68
	ResNet + CMT	97.98 ± 3.31	93.49 ± 6.52	92.02 ± 8.06	87.55 ± 15.86	96.49 ± 4.19
	ResNet + Time	95.97 ± 6.28	88.64 ± 8.16	89.24 ± 7.57	**89.94 ± 12.64**	88.56 ± 9.00
	ResNet + Time + C	97.17 ± 6.35	94.63 ± 5.84	**92.58 ± 8.06**	88.00 ± 16.06	97.16 ± 3.85
	ResNet + Time + CM	98.13 ± 3.49	94.55 ± 5.84	92.41 ± 8.07	87.56 ± 15.96	97.26 ± 3.81
	ResNet + Time + CT	98.20 ± 3.79	**94.69 ± 5.79**	92.55 ± 7.94	87.25 ± 15.89	**97.84 ± 3.28**
	Proposed Method	**98.54 ± 2.36**	94.58 ± 5.17	92.26 ± 7.74	87.25 ± 15.78	97.27 ± 3.42
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	ResNet + C	99.83 ± 0.81	98.63 ± 3.40	98.51 ± 3.30	98.50 ± 3.63	98.52 ± 5.59
	ResNet + CM	99.82 ± 0.74	98.70 ± 3.21	98.37 ± 3.56	97.83 ± 5.01	98.90 ± 4.94
	ResNet + CT	99.79 ± 0.92	98.54 ± 3.51	98.51 ± 3.13	98.54 ± 3.03	98.47 ± 5.69
	ResNet + CMT	99.86 ± 0.54	98.73 ± 3.18	98.42 ± 3.54	98.01 ± 4.90	98.84 ± 5.07
	ResNet + Time	99.66 ± 0.76	97.18 ± 3.50	96.51 ± 3.37	93.82 ± 6.46	99.20 ± 2.28
	ResNet + Time + C	99.88 ± 0.54	98.97 ± 2.22	98.79 ± 2.36	98.08 ± 4.19	99.51 ± 2.28
	ResNet + Time + CM	99.59 ± 1.52	98.91 ± 2.55	98.76 ± 2.51	97.92 ± 4.66	**99.60 ± 2.11**
	ResNet + Time + CT	**99.90 ± 0.47**	99.02 ± 2.01	98.90 ± 2.16	98.43 ± 3.49	99.36 ± 2.81
	Proposed Method	99.79 ± 0.82	**99.07 ± 2.12**	**98.96 ± 2.21**	**98.58 ± 3.48**	99.35 ± 2.97

The obtained results in the ablation study are presented in Table 6.4. In all the datasets, the temporal block enhances the performance of the methods with respect to the baseline *ResNet*. Similarly, it can be seen that the obtained results by the models with the context block are higher than the baselines (*ResNet* and *ResNet + Time*). In general, when time or motion is added to the context block, the models achieve better results. This means that the combination of visual, temporal, and contextual information produces a powerful discriminative model. It can also be observed that the higher performance obtained in

our model is an AUC value of 99.79% on the *Capri* dataset. On *Kvasir-Capsule* and *VH*, the obtained scores are 96.00% and 98.54%, respectively. Several reasons can justify the difference in performance among datasets, being the main differences between them: 1) the organ of study (colon on *Capri* vs. small bowel on *Kvasir-Capsule* and *VH*); 2) capsule device (Olympus EC-S10, Medtronic PillCam SB3, and Medtronic PillCam Colon2); and 3) amount of intestinal content. Therefore, capsule characteristics like optic, illumination, and resolution are not equivalent neither the intestinal mucosa and content. In addition, the statistics from each dataset are different as reported in Table 6.1. Despite all the mentioned differences, the results are coherent among the various datasets.

As previously stated, *Kvasir-Capsule* dataset lacks temporal information. To address this limitation, the frame index has been used as a substitute for temporal information. Despite this adjustment, similar effects on the system's performance have been observed in this dataset. This can be attributed to the fact that the order of the frames is a reliable proxy for timestamps.

The proposed method is compared with the following state-of-the-art methods: Zou et al. (2015), Chen et al. (2017a), Zhao et al. (2021), and Son et al. (2022). All these methods have been implemented, trained, and evaluated using the same datasets and evaluation methodology. The results reported in Table 6.5 show that the proposed method outperforms all others in all datasets.

Table 6.5: Comparison of the different methods of the state-of-the-art with our model in the image classification problem for each dataset. Displayed results are the mean obtained after evaluating a two-fold cross-validation.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	Zou et al. (2015)	75.37 ± 9.42	69.51 ± 11.67	69.11 ± 8.68	70.19 ± 16.35	68.03 ± 14.60
	Chen et al. (2017a)	83.65 ± 10.37	82.38 ± 9.20	76.90 ± 10.28	67.95 ± 16.18	85.84 ± 10.96
	Zhao et al. (2021)	94.05 ± 4.50	89.46 ± 7.72	85.09 ± 7.87	76.29 ± 14.40	93.89 ± 7.46
	Son et al. (2022)	95.75 ± 4.85	90.96 ± 6.56	81.03 ± 12.55	64.42 ± 26.40	$\mathbf{97.64 \pm 4.40}$
	Proposed Method	$\mathbf{96.00 \pm 4.57}$	$\mathbf{91.36 \pm 5.75}$	$\mathbf{87.47 \pm 7.49}$	$\mathbf{78.91 \pm 16.28}$	96.03 ± 4.29
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	Zou et al. (2015)	90.05 ± 9.91	84.56 ± 11.00	74.78 ± 12.22	56.87 ± 24.96	92.68 ± 9.98
	Chen et al. (2017a)	95.86 ± 5.46	90.29 ± 8.12	87.69 ± 8.28	82.53 ± 15.62	92.84 ± 10.02
	Zhao et al. (2021)	97.81 ± 4.24	93.56 ± 7.12	91.95 ± 8.04	87.54 ± 14.89	96.37 ± 5.07
	Son et al. (2022)	96.46 ± 6.65	89.27 ± 9.35	90.46 ± 8.73	$\mathbf{91.05 \pm 14.59}$	89.88 ± 9.21
	Proposed Method	$\mathbf{98.54 \pm 2.36}$	$\mathbf{94.58 \pm 5.17}$	$\mathbf{92.26 \pm 7.74}$	87.25 ± 15.78	$\mathbf{97.27 \pm 3.42}$
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	Zou et al. (2015)	86.06 ± 7.93	80.64 ± 12.24	65.93 ± 6.29	33.50 ± 12.51	98.35 ± 2.02
	Chen et al. (2017a)	95.28 ± 4.37	88.42 ± 7.84	88.69 ± 6.49	88.31 ± 10.25	89.07 ± 9.62
	Zhao et al. (2021)	99.85 ± 0.47	98.59 ± 2.23	98.17 ± 2.94	97.76 ± 3.74	98.58 ± 4.14
	Son et al. (2022)	$\mathbf{99.93 \pm 0.21}$	97.94 ± 2.74	96.06 ± 4.30	92.57 ± 8.22	$\mathbf{99.57 \pm 2.58}$
	Proposed Method	99.79 ± 0.82	$\mathbf{99.07 \pm 2.12}$	$\mathbf{98.96 \pm 2.21}$	$\mathbf{98.58 \pm 3.48}$	99.35 ± 2.97

6.6.2 Anatomical landmarks identification

In the second experiment, the difference between the predicted landmarks and the annotations provided by the experts is analyzed. On *Kvasir-Capsule* and *VH* datasets, the pylorus and the ileocecal valve, which delimit the small bowel, are identified. On the other hand, on *Capri* dataset, the boundaries of the colon, first cecal and last rectal images are used.

The results from the ablation study come from minimizing the rectangular pulse function over the output signal of each setting. Table 6.6 contains the MAE and median error of each one in frames and time. The reported results show that the use of the temporal and context block reduces the error of the baseline *ResNet*. Particularly, in the small bowel datasets, *Kvasir-Capsule* and *VH*, there is a large difference between MAE and median error. This suggests that there are several outliers. Despite them, the proposed method achieves promising results in all the cases.

Table 6.6: Comparison of the ablation study in the anatomical landmarks identification task for each dataset. MAE and median error are represented as the difference in frames and time (hh:mm:ss).

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	ResNet	668.34 ± 1091.84	-	111.75 ± 28.25	-	1875.96 ± 2747.54	-	1124.00 ± 493.00	-
	ResNet + C	1505.83 ± 3316.56	-	207.50 ± 153.00	-	2147.42 ± 2967.07	-	908.50 ± 757.50	-
	ResNet + CM	1504.54 ± 3272.05	-	217.25 ± 131.75	-	1877.62 ± 2736.03	-	928.00 ± 684.50	-
	ResNet + CT	1677.79 ± 3509.05	-	139.00 ± 51.50	-	1902.29 ± 2683.68	-	1220.50 ± 391.50	-
	ResNet + CMT	830.08 ± 1341.51	-	127.50 ± 31.00	-	1770.79 ± 2771.27	-	743.75 ± 580.25	-
	ResNet + Time	785.88 ± 1182.88	-	93.00 ± 15.50	-	2002.38 ± 2959.43	-	1077.00 ± 539.00	-
	ResNet + Time + C	535.50 ± 1089.54	-	**26.75 ± 2.25**	-	1710.71 ± 2769.83	-	559.00 ± 169.00	-
	ResNet + Time + CM	606.08 ± 1104.42	-	61.25 ± 9.25	-	1727.67 ± 2767.06	-	651.50 ± 383.00	-
	ResNet + Time + CT	556.38 ± 951.79	-	116.00 ± 43.00	-	1730.50 ± 2756.90	-	663.25 ± 83.25	-
	Proposed Method	**465.88 ± 918.13**	-	76.50 ± 46.50	-	**1679.67 ± 2775.72**	-	**487.25 ± 163.75**	-
VH	ResNet	667.70 ± 1070.13	00 : 04 : 00	220.75 ± 191.75	00 : 01 : 45	2710.98 ± 4505.58	00 : 22 : 02	1235.75 ± 1126.75	00 : 12 : 20
	ResNet + C	559.98 ± 1007.92	00 : 03 : 11	78.50 ± 17.00	00 : 00 : 34	1290.33 ± 2117.18	00 : 14 : 52	198.50 ± 82.00	00 : 04 : 02
	ResNet + CM	512.38 ± 870.20	00 : 02 : 59	91.75 ± 48.75	00 : 00 : 39	1028.02 ± 1704.22	00 : 12 : 14	199.50 ± 85.50	00 : 03 : 24
	ResNet + CT	557.83 ± 1181.40	00 : 03 : 10	78.25 ± 16.75	00 : 00 : 32	1279.73 ± 1774.48	00 : 14 : 55	369.75 ± 3.75	00 : 05 : 53
	ResNet + CMT	500.90 ± 1166.62	00 : 02 : 45	50.50 ± 17.0	00 : 00 : 25	1077.35 ± 1605.62	00 : 12 : 05	259.50 ± 109.00	00 : 05 : 40
	ResNet + Time	731.71 ± 1451.30	00 : 03 : 57	103.50 ± 59.50	00 : 01 : 14	2050.94 ± 4142.68	00 : 17 : 01	397.00 ± 350.00	00 : 06 : 26
	ResNet + Time + C	502.62 ± 877.55	00 : 02 : 47	59.50 ± 4.00	00 : 00 : 25	911.23 ± 1619.69	00 : 11 : 38	167.25 ± 106.25	00 : 03 : 03
	ResNet + Time + CM	**417.79 ± 814.67**	**00 : 02 : 18**	44.00 ± 7.00	**00 : 00 : 10**	1110.48 ± 1992.44	00 : 13 : 02	**166.50 ± 103.00**	**00 : 03 : 00**
	ResNet + Time + CT	443.46 ± 1092.63	00 : 02 : 31	50.25 ± 8.25	00 : 00 : 23	1051.17 ± 1933.66	00 : 11 : 37	225.25 ± 170.25	00 : 03 : 29
	Proposed Method	443.69 ± 1064.05	00 : 02 : 38	**41.50 ± 11.00**	00 : 00 : 15	**837.77 ± 1485.79**	**00 : 09 : 46**	210.75 ± 164.75	00 : 03 : 14
Capri	ResNet	53.70 ± 110.44	00 : 01 : 19	14.50 ± 4.00	00 : 00 : 06	13.80 ± 48.35	00 : 02 : 19	1.00 ± 0.00	**00 : 00 : 01**
	ResNet + C	28.94 ± 85.33	00 : 00 : 59	4.50 ± 0.50	00 : 00 : 02	41.76 ± 215.63	00 : 04 : 44	3.00 ± 0.00	**00 : 00 : 01**
	ResNet + CM	32.43 ± 97.72	00 : 01 : 00	2.75 ± 0.25	**00 : 00 : 01**	38.58 ± 210.08	00 : 03 : 01	1.75 ± 0.25	**00 : 00 : 01**
	ResNet + CT	32.62 ± 91.01	00 : 01 : 02	5.00 ± 1.00	00 : 00 : 02	48.45 ± 240.70	00 : 07 : 08	2.00 ± 0.50	00 : 00 : 02
	ResNet + CMT	32.35 ± 100.58	00 : 01 : 01	2.50 ± 0.50	**00 : 00 : 01**	37.71 ± 209.38	00 : 03 : 06	2.00 ± 0.00	**00 : 00 : 01**
	ResNet + Time	38.40 ± 86.46	00 : 01 : 11	5.00 ± 2.00	00 : 00 : 02	19.80 ± 114.86	00 : 05 : 37	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + C	29.57 ± 81.18	00 : 00 : 51	4.50 ± 1.50	**00 : 00 : 01**	8.89 ± 38.02	00 : 02 : 13	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CM	**23.58 ± 69.27**	**00 : 00 : 41**	3.50 ± 0.50	**00 : 00 : 01**	8.26 ± 38.02	00 : 02 : 04	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CT	30.40 ± 79.19	00 : 00 : 58	5.25 ± 3.25	**00 : 00 : 01**	8.88 ± 38.15	00 : 02 : 09	**1.00 ± 0.00**	**00 : 00 : 01**
	Proposed Method	29.40 ± 83.94	00 : 00 : 55	**2.75 ± 0.25**	**00 : 00 : 01**	**7.91 ± 37.82**	**00 : 01 : 47**	**1.00 ± 0.00**	**00 : 00 : 01**

The proposed method is compared with Zhao et al. (2021) and Son et al. (2022), as shown in Table 6.7. For this experiment, the proposed methods by Zou et al. (2015) and Chen et al. (2017a) have not been considered since they do not identify the boundaries of the organs. The results in Table 6.7 show that the proposed method outperformed all methods in all the cases.

It can be observed in Table 6.6 and Table 6.7 that in both small bowel datasets, that

Table 6.7: Comparison of the different methods of the state-of-the-art with our model in the anatomical landmarks identification task for each dataset. MAE and median error are represented as the difference in frames and time (hh:mm:ss).

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	Zhao et al. (2021)	2644.16 ± 4637.65	-	1251.00 ± 115.25	-	4603.58 ± 1545.95	-	1669.00 ± 185.26	-
	Son et al. (2022)	2711.00 ± 3435.83	-	1786.00 ± 231.93	-	2409.00 ± 3106.30	-	1506.75 ± 1161.42	-
	Proposed Method	$\mathbf{465.88 \pm 918.13}$	-	$\mathbf{76.50 \pm 46.50}$	-	$\mathbf{1679.67 \pm 2775.72}$	-	$\mathbf{487.25 \pm 163.75}$	-
VH	Zhao et al. (2021)	1304.23 ± 1394.26	00 : 08 : 20	915.25 ± 22.98	00 : 04 : 34	3308.58 ± 583.28	00 : 31 : 44	1765.25 ± 461.03	00 : 16 : 59
	Son et al. (2022)	1390.47 ± 3487.30	00 : 07 : 12	304.75 ± 220.97	00 : 02 : 14	1552.00 ± 520.78	00 : 16 : 47	627.75 ± 1469.01	00 : 08 : 09
	Proposed Method	$\mathbf{443.69 \pm 1064.05}$	**00 : 02 : 38**	$\mathbf{41.50 \pm 11.00}$	**00 : 00 : 15**	$\mathbf{837.77 \pm 1485.79}$	**00 : 09 : 46**	$\mathbf{210.75 \pm 164.75}$	**00 : 03 : 14**
Capri	Zhao et al. (2021)	214.24 ± 437.6	00 : 07 : 28	85.0 ± 23.33	00 : 01 : 11	524.94 ± 1258.34	00 : 35 : 23	23.5 ± 2.12	00 : 01 : 02
	Son et al. (2022)	56.39 ± 161.94	00 : 04 : 00	14.50 ± 0.70	00 : 00 : 08	32.25 ± 111.27	00 : 07 : 29	7.50 ± 0.70	00 : 00 : 43
	Proposed Method	$\mathbf{29.40 \pm 83.94}$	**00 : 00 : 55**	$\mathbf{2.75 \pm 0.25}$	**00 : 00 : 01**	$\mathbf{7.91 \pm 37.82}$	**00 : 01 : 47**	$\mathbf{1.00 \pm 0.00}$	**00 : 00 : 01**

the identification of the entrance of the organ is more accurate than the exit. On the other hand, for the colon dataset (*Capri*) the identification of last rectal image is more accurate than the entrance.

Finally, the impact of the landmarks identification strategy is analyzed. To evaluate it, the third step of our model has been applied to the outputs of the models from Zhao et al. (2021) and Son et al. (2022). Table 6.8 summarizes the results obtained, which indicate that by solving the minimization problem the errors decrease. These findings suggest that the proposed strategy of locating anatomical landmarks is more effective than those currently published.

Table 6.8: Comparison of the different strategies for identifying the anatomical landmarks. The proposed strategy is applied to state-of-art methods. The displayed results are the median error obtained after evaluating a two-fold cross-validation.

Method	Dataset					
	Kvasir-Capsule		VH		Capri	
	Entrance	Exit	Entrance	Exit	Entrance	Exit
Zhao et al. (2021)	1251.00	1669.00	915.25	1765.25	85.00	23.50
Zhao et al. (2021) + Step 3	93.00	702.50	65.00	478.00	3.50	1.00
Son et al. (2022)	1786.00	1506.00	304.75	627.75	14.50	7.50
Son et al. (2022) + Step 3	686.25	1683.25	214.50	1236.00	35.50	6.00
Proposed Method	**76.50**	**487.25**	**41.50**	**210.75**	**2.75**	**1.00**

The results from both tasks, image classification (Table 6.5) and anatomical landmark identification (Table 6.7) show a strong correlation. In other words, the better the classifier, the more accurate the limit identification of the organs.

6.6.3 Qualitative Results

This section aims to gain additional information about the performance of the proposed method and to visualize the types of errors in a qualitative manner (Figure 6.6). These results are presented using one test video per dataset (*Kvasir-Capsule* dataset in Figure 6.6a, *VH* dataset in Figure 6.6b and *Capri* dataset in Figure 6.6c). The selected ones have metrics near the median values reported in Table 6.6, thus avoiding outliers and championship cases. The output signals represent the classification task, where the probability for each image to belong to the studied organ is plotted. Yellow dots are the frames corresponding to the organs (small bowel or large intestine), whereas the blue ones are considered out of this range. The visualization allows assessing the performance of our method by removing certain blocks and understanding the contribution of each component to the overall system. For that purpose, each figure shows four of the methods previously introduced: *ResNet*, *ResNet + Time*, *ResNet + CMT* and *Proposed Method*.

For the *Kvasir-Capsule* dataset (Figure 6.6a), the *ResNet + Time* method infers worse probabilities than the *ResNet* model. When the temporal information is combined with the contextual block in our method, the obtained output signal is smoother and more similar to the ground truth. However, there are still some misclassified sequences outside the small intestine. In the case of the videos from the other datasets (Figure 6.6b and Figure 6.6c) the application of the contextual information further evidence the improved performance of the proposed method even more.

Moreover, each subfigure contains a set of false positive and false negatives samples determined by our method. The small bowel videos show several misclassified examples, where the mucosa is completely hidden, thus preventing the system from making a correct prediction.

The identification of the anatomical landmarks is a complex task since only one correct frame in the video has to be determined as the entrance or exit of the organ. In Figure 6.6, the green dashed lines over the output signals indicate the frame labeled by the experts while the purple ones correspond to the system predictions. The rectangular pulse function fits the output signal and correctly identifies the landmarks, but it fails when the prediction of belonging to the organ is wrong. In the identified last pylorus image, it can be observed that despite the distance between the predicted and the real landmarks, the frames are visually similar (Figure 6.6a and 6.6b). But when the mucosa is hidden by the noise content of the GI tract, as happens in the exit sequences in Figures 6.6a and 6.6b, the error is higher. Therefore, the complexity of the problem increases. However, in *Capri* dataset (Figure 6.6c), the exit of the large intestine is easier to identify due to the drastic change in the visual features caused by the evacuation of the capsule from the body or because the video stopped.

6.7 Discussion and Conclusion

In this paper, an effective deep learning system for WCE is proposed. The method is designed to, firstly, infer the probability for every image to belong to the area of interest, taking advantage of temporal, neighboring and motion information. Secondly, the landmarks are predicted by solving a minimization problem. Experimental results have been reported over three datasets, one public and two private. In all of them, the proposed method improves the results of the baseline system.

The results obtained in the two datasets of the small bowel, *Kvasir-Capsule* and *VH*, show a high performance in the classification problem. Moreover, our method in the *VH* dataset displays even better results, increasing at least three points in each metric. Several reasons can justify the difference in performance achieved on each dataset such as dataset size, amount of intestinal content or device used to collect the video.

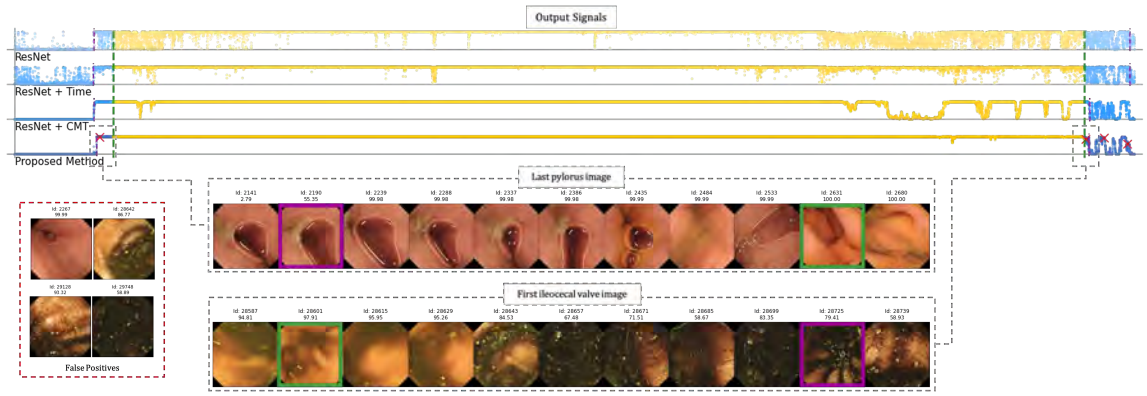
After performing all the experiments and analyzing the results, we believe that our method is a good candidate for the automatic classification of organs regardless of the device used. Although our method of landmark identification does not achieve the best performance for all the datasets, it exhibits promising results.

One limitation of the proposed system is that it has been designed to deal with only one organ, given the lack of multi-organ labels in the used datasets. However, this limitation could be addressed by incorporating anatomical landmarks for multiple organs and adapting the CMT block accordingly. Additionally, the results strongly depend on the organ, the used WCE device, and the dataset size. To overcome this issue, future research could focus on exploring a general method for multiple devices and organs.

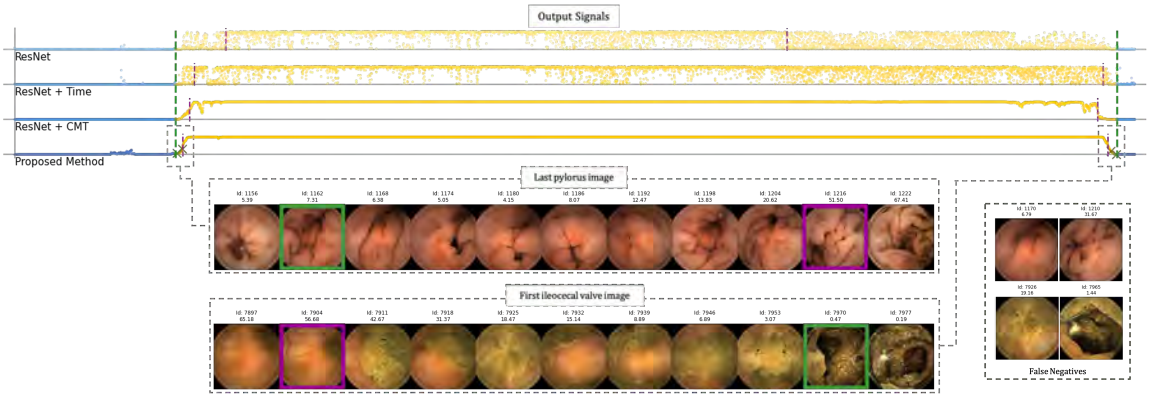
Future work could explore the detection of multiple organs and their anatomical landmarks. Moreover, distinct landmarks inside one organ can be localized, for example, the flexures of the large intestine.

Acknowledgment

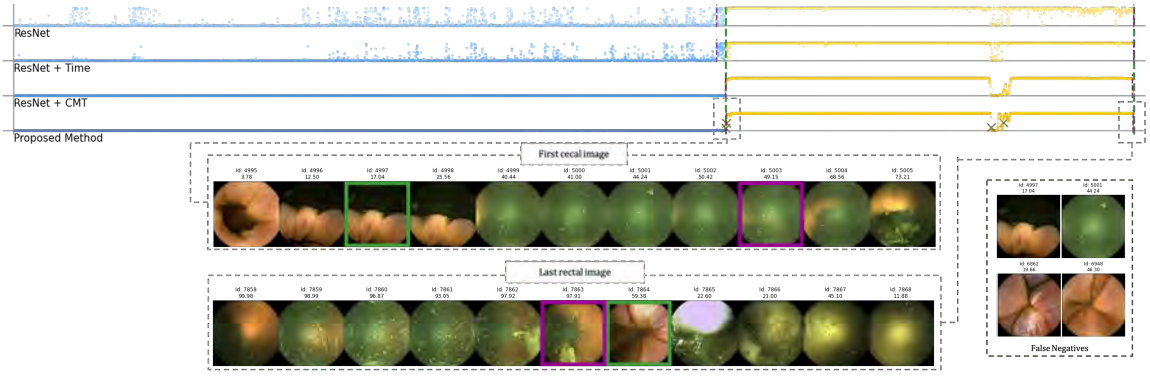
This work has been partially funded by MINECO Grant RTI2018-095232-B-C21, SGR 1742 and 01104 (Generalitat de Catalunya), Innovate UK project 104633, AI AWARD02440 project from NIHR and NHSX and the FPU grants FPU17/02727 and FPU20/01090 (Ministerio de Universidades, Spain). The authors would also like to thank NVIDIA for their GPU donations.



(a) Kvasir-Capsule



(b) VH



(c) Capri

Figure 6.6: Visual representation of the system outputs of three WCE videos from: a) *Kvasir-Capsule* b) *VH* and c) *Capri* datasets. Each subfigure contains the output signals and the identification of the anatomical landmarks for the evaluated methods. Yellow points represent frames from the organ of interest (small bowel or large intestine), whereas the blues ones are outside these areas. The second task is displayed over the outputs signals as dashed lines. The predicted landmarks are ticked in purple, while the ground truth is in green. Below the output signals are displayed a uniform sampling of frames around the landmarks, achieving sequences of 11 items. The frame identification (id) and the probability of belonging to the organ of interest are shown above each image. The frames of the labeled and predicted landmarks are surrounded by a green and purple box, respectively. Finally, several misclassified frames are shown, which are localized in the output signal of the *Proposed Method* as crosses in red for false positives and dark green for false negatives samples. The figure is best viewed on the computer.

Chapter 7

Paper IV

Using the triplet loss for domain adaptation in WCE

Pablo Laiz[a], Jordi Vitrià[a], and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) 399-405, 10 2019.

doi: 10.1109/ICCVW.2019.00051

Contents

Motivation and Context

The evolution of hardware allows the design of smaller and more powerful parts of the WCE, leading to the development of new capsule devices. The comparison between images captured by distinct pill cameras enables us to perceive that features such as illumination and resolution are different. Therefore, this results in datasets from different capsules having significant changes in data distribution.

As previously mentioned in Chapter 3, DL exhibit optimal performance when the training and test data distributions are identical. However, when they are obtained from different distributions the performance of the system deteriorates. This is known as covariate distribution shifts. In the context of this thesis, this phenomenon arises when a dataset is evaluated on a DL system trained with data from another capsule device. Consequently, the systems may no longer work well in images captured by new capsules.

The labeling of an entirely new dataset for each new capsule can be a time-consuming and expensive process. Therefore, discarding previous databases and starting from scratch each time a new device is released is impractical. To address this covariate distribution shift, a domain adaptation approach is used. The following study proposed the use of the TL to improve the generalization of the model over several datasets from different devices, particularly PillCam SB2 and PillCam SB3. The method is categorized as a deep, supervised, single-modality, and one-step approach according to Section 3.3.1.

The proposed method works by adapting an embedding space that is trained with a large dataset to a new domain where only a comparatively small labeled dataset is available. The space is adapted by generating triplets of images from both domains, so two images of the same category but from different domains are closer to each other than images belonging to the same domain but in a different class. This approach aligns the embedding space without learning device-specific features while maintaining the accuracy of the model in both domains.

The experimental results show that the designed method performs properly with both capsule devices, proving its ability to adapt the embedding space to new domains. The quantitative and qualitative results suggest that this approach is a valuable tool for researchers to create new models that are effective across different datasets and devices. Therefore, there is no need to spend significant time and resources to create new large databases for each new device.

Using the triplet loss for domain adaptation in WCE

Pablo Laiz[a], Jordi Vitrià[a], and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

7.1 Abstract

Wireless Capsule Endoscopy (WCE) is a minimally-invasive procedure that, based on a vitamin-size camera that is swallowed by the patient, allows the visualization of the entire gastrointestinal tract. This technology was developed 20 years ago to perform useful and safe studies of different bowel disorders. However, especially the number of captured images and their difficult interpretation has hindered its deployment in some clinical scenarios.

Deep learning methods have the necessary capacity to deal with WCE image interpretation, but training good models is still an open problem for some bowel disorders due to the fact that obtaining a sufficiently large set of positive cases, for the creation and validation of the model, is an arduous task. Moreover, technological advances are rapidly moving forward proposing new hardware able to obtain images with a substantially improved quality. Given these two facts, it is obvious that highly accurate models can only be built by considering heterogeneous datasets composed of images captured by different cameras, and if training methods are able to find invariances with respect to the image acquisition systems.

In this paper, we study the use of deep metric learning, based on the triplet loss function, to improve the generalization of a model over different datasets from different versions of WCE hardware. The obtained results show evidence that with just a few labeled images from a newer camera set, a model that has been trained with images from older systems can be easily adapted to the new environment.

Key words: deep learning; deep metric learning; triplet loss; domain adaptation; capsule endoscopy; medical imaging.

The paper has been re-typeset to match the thesis style.

7.2 Introduction

Wireless Capsule Endoscopy (WCE) is a medical procedure that enables the visualization of the entire gastrointestinal tract. WCE is based on a vitamin-size capsule, equipped with a light source, camera, an optical lens, radio transmitter, and a battery, that is swallowed by the patient and propelled by the peristalsis along all Gastrointestinal (GI) tract, allowing the full visualization of it, from inside, without pain or sedation.

The use of a WCE capsule produces a long video that contains thousands of images that must be individually reviewed by a medical specialist, making the interpretation and analysis of WCE data a complex and time-consuming activity. To overcome this drawback we can use a Computer-Aided Detection and Diagnosis (CAD) to support human interpretation.

The first difficulty that researchers must tackle when developing CADs for WCE is the need to build representative databases for some specific disease or condition. The creation of these databases is time-consuming and economically expensive because of technical questions and also because of the scarcity of positive cases. For this reason, most of the methods we can find in the literature are built and validated with very small datasets.

Another important point that should not be overlooked is that, in the medical field, technological advances are rapidly moving forward. Since the presentation of the first WCE device in 2001, new devices have been periodically presented with better image resolution, illumination or larger field of view. Today, we can find different WCE devices, coming from different manufacturers, that present different technical specifications. Table 7.1 illustrates some of the most known WCE devices with their main specs, and Figure 7.1 shows images captured by two different capsules from Medtronic: PillCam SB2 and PillCam SB3. As it can be appreciated, images from PillCam SB3 are better. It is clear that if a model is trained with data from an older capsule, it might not give the expected results when it is evaluated on a newer one since the same data distribution is not guaranteed. However, when the cost of creating a database is that high, it is not acceptable to lose previous databases and build a new one from scratch each time a new device is released.

Table 7.1: Capsule endoscopy devices used to perform endoscopy operations. The table contains a summary of the main features of each one.

Capsule	**Size (mm)**	**Weight (g)**	**Battery life (h)**	**Resolution (pixels)**	**Frames per second**	**Field of view**
PillCam SB2 - Given Imaging	26.0 × 11.0	3.40	8	256×256	2 fps	156°
PillCam SB3 - Given Imaging	26.2 × 11.4	3.00	>8	340×340	2-6 fps	156°
EndoCapsule - Olympus America	26.0 × 11.0	3.50	>8	512×512	2 fps	145°
MiroCam - IntroMedic Company	24.5 × 10.8	3.25 - 4.70	>11	320×320	3 fps	170°
OMOM Jinshan - Science and Technology	27.9 × 13.0	6.00	>6 - 8	640×480	2 fps	140°

To overcome this problem, we propose a domain adaptation method based on deep metric learning using the triplet loss. The proposed method aims to adapt the embedding

Figure 7.1: Frames from different capsules present different technical quality. (a) PillCam SB2 capsule image (b) PillCam SB3 capsule image. Image in (b) is clearly better than image in (a).

space trained with a large training data set to a new domain where only comparatively few labeled images are available. The embedding space is adapted by generating triplets of images from both domains, with the goal that two images in the same category are closer than images belonging to different domains. The obtained results show that by using a small labeled dataset from the new domain, the embedding space can be adapted to work **in both domains** with high performance.

The rest of the paper is organized as follows: first, we give an overview of the related work in the field. This is followed by a description of our methodology, presenting the system architecture, followed by the experimental setup and results. Finally, we conclude the paper and give directions for future work.

7.3 Related Work

7.3.1 Deep Learning for WCE analysis.

Several deep learning methods have been proposed for WCE image analysis, dealing with different pathologies such as bleeding, hemorrhages, angiectasia, polyps/cancer, ulcers, and hookworms. For example, Zou et al. (2015) proposed a CNN-based method to classify the different organs of the digestive system such as the stomach, small intestine, and colon; Seguí et al. (2016) proposed a classification method of motility events such as turbid, bubbles, clear blob, wrinkle, and wall; finally, Yuan and Meng (2017) proposed a stacked sparse autoencoder-based approach for detecting polyps.

7.3.2 Metric Learning

Metric learning has been extensively used in many machine learning and computer vision applications (Kulis, 2013). Inspired by the success of deep neural networks, deep metric

learning has become popular in the last few years. These methods aim to learn a discriminative feature embedding, using deep neural networks, such that similar samples are represented by similar embedding vectors and different samples are represented by dissimilar ones. In order to learn these features, embedding deep neural networks are trained using special loss functions such as the Contrastive loss (Hadsell et al., 2006), the Triplet loss (Hoffer and Ailon, 2015) or the Quadruplet loss (Chen et al., 2017b). Triplet loss has shown very good results on several image retrieval tasks (Chechik et al., 2010; Gordo et al., 2016) and in many image classification problems such as facial recognition (Schroff et al., 2015), person re-identification (Cheng et al., 2016; Hermans et al., 2017) or action recognition (Ma et al., 2016).

The selection of the triplets is one of the key factors when implementing the triplet loss. In the literature, we can find several methodologies, such as the *Batch All* or *Batch Hard* (Ding et al., 2015), that face the problem of triplet sampling for each batch.

7.3.3 Domain Adaptation

Domain adaptation methods are designed to deal with the problem of distribution shift across domains. Many domain adaptation (or transfer learning) approaches have been proposed for computer vision applications (Oquab et al., 2014; Donahue et al., 2014; Chen et al., 2015). To our knowledge, the use of triplet loss in the domain adaptation problem has been limited. Huang et al. (2015) defined a triplet visual similarity constraint for learning to rank across two sub-networks using online and offline images.Yu et al. (2018) used the triplet loss to correct the selection bias in the triplet selection. Deng et al. (2018) used the triplet loss and pseudo-labels for unsupervised domain alignment.

7.4 Method

The architecture of our system is illustrated in Figure 7.2. As it can be seen, the system architecture consists of a classical neural network architecture followed by a normalization layer L_2 and an embedding layer which is optimized with the triplet loss.

In this section, we first introduce the triplet loss function for deep metric learning and then we consider its role in the problem of domain adaptation in our scenario.

7.4.1 Triplet loss for Deep Metric Learning

Let X, Y denote two random variables, which indicate data and label, respectively. Let D be the set of data sampled from $P(X, Y)$. The goal of metric learning is to learn a distance function that assigns small (or large) distance values to a pair of similar (or dissimilar) images. Deep metric learning uses a deep neural network to learn a feature embedding

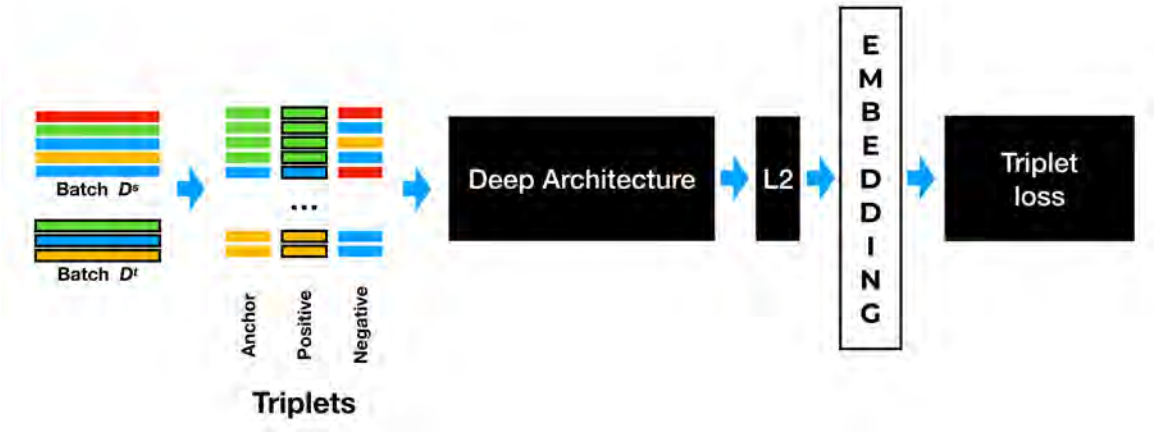


Figure 7.2: Overview of the proposed CNN structure. The input of the network consists of a batch of images from both domains, source D^s and target D^t. A set of triplets is generated, where anchor images and negative images are from the same domain while positive images are from a different domain but belonging to the same class as the anchor image. The architecture is defined as a standard CNN Architecture followed by L_2 normalization and an embedding layer. It is optimized by using the triplet loss over the generated triplets.

$x' = \Phi(x)$ with the goal of learning a non-linear distance function as follows:

$$d^2(x^i, x^j) = ||\Phi(x^i) - \Phi(x^j)||_2^2 \tag{7.1}$$

In order to learn this embedding representation $\Phi(x_i)$, the triplet loss function is defined as follows:

$$L_{triplet} = \sum_{(x^a, x^p, x^n) \in D} \Big[d^2(x^a, x^p) - d^2(x^a, x^n) + \alpha \Big]_+ \tag{7.2}$$

where $[\cdot]_+ = max(\cdot, 0)$, $\alpha > 0$ and x^a, x^p and x^n refers to anchor, positive and negative examples respectively. Hence, the set of triplets used to train the network is defined as:

$$\tau = \{(x^a, x^p, x^n) | y^a = y^p \text{ and } y^a \neq y^n\} \tag{7.3}$$

This loss function has shown excellent results in learning feature embedding mappings, requiring that the distance between $\Phi(x_a)$ and $\Phi(x_p)$ is smaller than the distance between $\Phi(x_a)$ and $\Phi(x_n)$.

The selection of triplets during training is one of the key factors in order to optimize the network using the triplet loss. As it was said before, there exist two main methodologies to face the sampling problem of triplets: *Batch All* and *Batch Hard* (Ding et al., 2015). In *Batch All* strategy a batch of images from the training set is selected and then all possible triplets are generated to optimize the loss. On the other hand, in *Batch Hard* strategy, triplets are generated by seeking, for each sample x_a in the batch, the hardest positive sample, or farthest positive sample $argmax_{x^p}(\|\Phi(x^a) - \Phi(x^p)\|_2^2)$, and the hardest

negative sample, or closest negative sample $argmin_{x^n}(\|\Phi(x^a) - \Phi(x^n)\|_2^2)$. Depending on the data set, *Batch All* can lead to a sub-optimal solution while *Batch Hard* can have some convergence problems as a consequence of only considering the hardest samples. For our problem, we will consider the *Batch All* strategy due to the visual heterogeneity of our classes.

7.4.2 Domain adaption using triplet loss

In our problem, it is assumed that two datasets from different domains are available, the source domain dataset, D^s, and the target domain dataset, D^t, obtained by different capsules. Both datasets are fully labeled but D^s is expected to contain a larger amount of images while D^t is expected to be smaller. To goal is to adapt the model trained with images from D^s to the new environment D^t with minimal effort.

We assume that there is a covariate shift on the marginal distribution $P(X)$ across domains while the conditional distribution $P(Y|X)$ remains equal. To correct the distribution shift across domains, we first learn a model that defines the embedding function using the large labeled training set D^s. This model is trained using the standard *Batch All* strategy using all the images from the training set D^s. Then, in order to align the data distributions from both domains and then reduce the whole distribution discrepancy between the source and target datasets, new triplets are generated using both domains, D^s and D^t. Triplets are generated from a batch of N images, where K images are selected from D^t while $N-K$ from D^s. Each triplet consists of an anchor sample x^a that can be from D^s or D^t indifferently, a positive sample x^p that is from a different domain than x^a but with the same label and a negative sample x^n which is from the same domain than the anchor image x^a. Formalizing, the set of triplets used to train the system is defined as follows:

$$\tau = \{(x_i^a, x_j^p, x_i^n) | y_i^a = y_j^p \text{ and } y_j^a \neq y_j^n \text{ and } i \neq j\}$$

where i and j represent any of the classes of the dataset.

7.5 Experimental Results

7.5.1 Dataset

In order to validate the proposed system, two different datasets have been used, named **SB2D** and **SB3D**. These datasets have been created using two different versions of the capsules. $SB2D$ has been created using the PillCam SB2 version while the $SB3D$ dataset with the PillCam SB3 version. The most remarkable difference between these two capsules is a 30% improvement in resolution quality (see Table 7.1 and Figure 7.1) but also the improvement in illumination, color and the overall image quality.

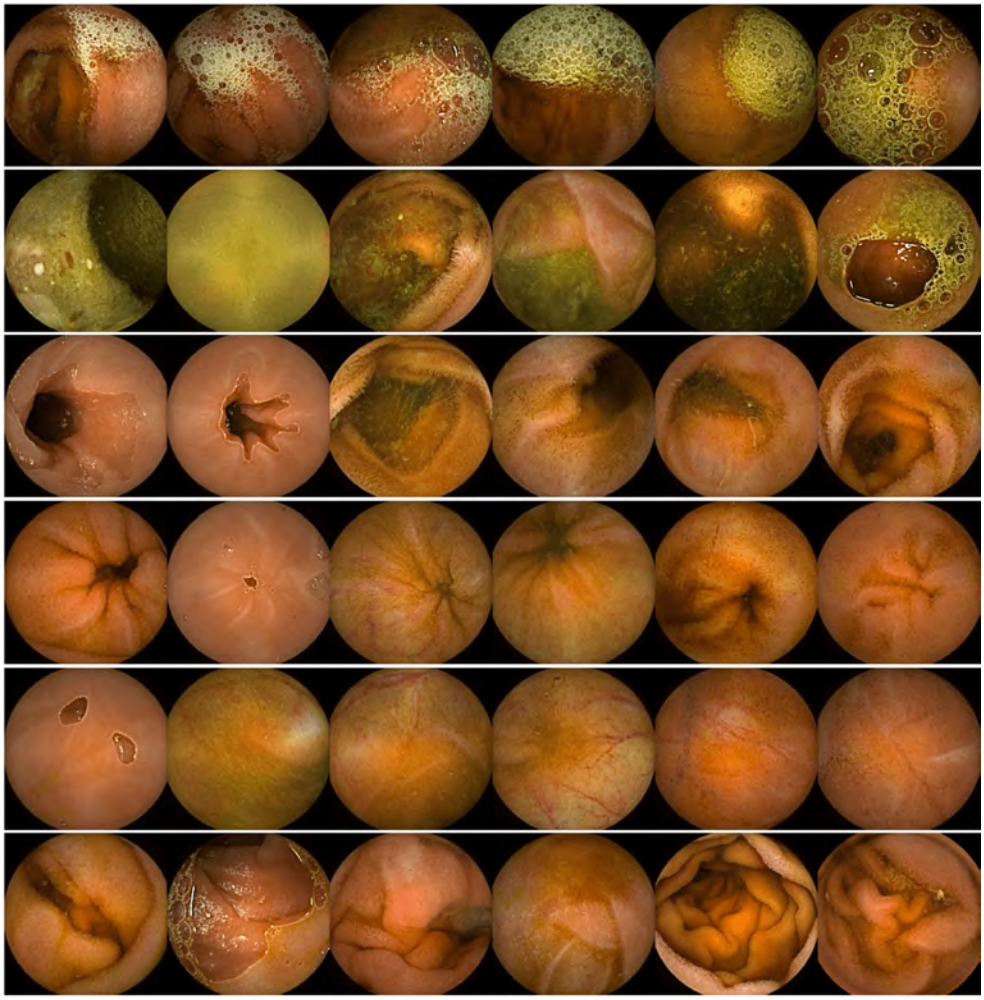


Figure 7.3: Each row shows six exemplary images for each category in the database: *bubbles, turbid, clear blob, wrinkles, wall*, and *undefined*, respectively

Both datasets were labeled by expert physicians into 6 different classes: *bubbles, turbid, clear blob, wrinkles, wall*, and *undefined*. All images are resized to 256×256 pixels. $SB2D$ contains a total of 120K labeled images, 20K images per class from a total of 50 different procedures. $SB3D$ contains a total of 6K images, 1K images per class obtained from a total of 10 different procedures. Figure 7.3 shows six exemplary images for each class.

7.5.2 Implementation Details and Evaluation Methodology

We implemented the methods using Tensorflow (Abadi et al., 2015). The system architecture is based on the ResNet-50 (He et al., 2016) with an additional normalization L_2 layer and embedding layer of size 2048. ResNet parameters are preloaded from a trained network using Imagenet dataset. The network is trained for a total of 50 epochs using the stochastic gradient descent (SGD) algorithm with a cyclic learning (Smith, 2017) rate that moves

between 0.01 and 1e-5 with stepsize 4000. The batch size is set to 64. All experiments are executed using the standard 2-fold cross-validation methodology where images of the same procedure strictly belong to only one partition.

7.5.3 Results

A first experiment is done to compare the proposed methodology, **TL_SB2-3**, against 3 classical training alternatives **CE_SB2**, **CE_SB2-FT-SB3** and **TL_SB2**. CE_SB2 refers to ResNet-50 trained on $SB2D$ with the classical cross-entropy loss function. CE_SB2-FT-SB3 consists of the CE_SB2 model where the classification layer is fine-tunned using the standard methodology with the dataset $SB3D$. TL_SB2 refers to the proposed architecture presented in Figure 7.2, based on the ResNet-50 and optimized with the triplet loss function with the dataset $SB2D$. Finally, the proposed method TL_SB2-3 which is optimized with the triplet loss function using data from both domains, $SB2D$ and $SB3D$. In order to avoid overfitting, the parameters of the network are initialized using the TL_SB2 model which is trained using $SB2D$.

As it can be seen in Table 7.2, CE_SB2 and TL_SB2 obtain good results on $SB2D$ but very poor results when using $SB3D$. On the other hand, CE_SB2-FT-SB3, that uses the classical fine-tunning procedure, obtains satisfactory results on $SB3D$ but its accuracy on $SB2D$ drops. The proposed methods, TL_SB2-3 is able to obtain good results on $SB3D$ without deteriorating its accuracy on the source domain $SB2D$.

Table 7.2: Comparison of the different proposed methods evaluated in target and source domains respectively, $SB2D$ and $SB3D$.

Methods	**Accuracy (%)**	
	SB2	**SB3**
CE_SB2	92.5	51.7
CE_SB2-FT-SB3	62.7	87.0
TL_SB2	93.3	41.2
TL_SB2-3	93.1	89.3

Figure 7.4 shows the UMAP (McInnes et al., 2018) plots of the learned embedding spaces. Each color represents a different class. Plot (a) illustrates the embedding space obtained with $SB2D$; in plot (b) colored points represent $SB3D$ data projected into the $SB2D$ embedding space (gray); and plot (c) illustrates the adapted embedding space with $SB2D$ (grey) and $SB3D$ (colored). As it can be observed, there exists a clear shift between the distribution from different domains which is adapted after training with both domains.

In the second experiment (see Table 7.3), we evaluated the accuracy of the system using different amount of images per procedure. A total of 10 procedures were selected using the standard 2-fold cross-validation strategy. As it can be seen, with just 30 images per

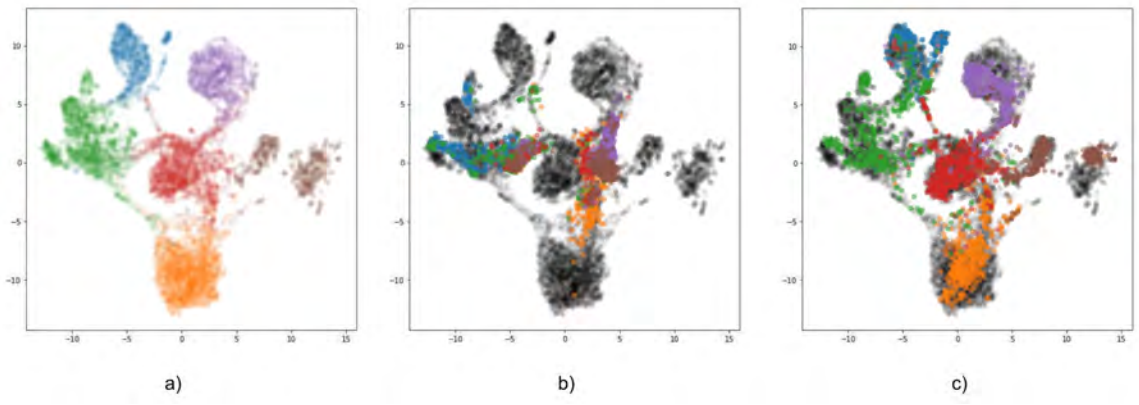


Figure 7.4: UMAP plots of the learned embedding spaces. Each color represent a different class while (a) illustrates the embedding space obtained with $SB2D$; (b) colored points represents $SB3D$ data projected into the $SB2D$ embedding space (gray); and (c) illustrates the adapted embedding space with $SB2D$ (gray) and $SB3D$ (colored).

procedure (5 images per class), i.e. a total of 150 images since 5 videos are used for creating the training data, the accuracy of the system is increased from 41.28% to 84.64%. As more images per procedure are used, the accuracy increases, obtaining an accuracy of 89.32% when all images of all procedures are used.

Table 7.3: Accuracy of the proposed system evaluated on $SB3D$ with different size of training samples from the target domain. Data is obtained uniformly per class (k = 6) and procedure (n = 5).

Method	**SB3 Images**	**Accuracy** (%)
TL_SB2-3	0	41.2
	150	84.6
	300	86.1
	750	87.3
	1500	88.6
	3000	89.3

Finally, Table 7.4 shows the behavior of the system when more diversity of the target domain is introduced. To perform this experiment, the accuracy of the system is evaluated when a different amount of videos are used but setting the same amount of labeled data, 600 images. As it can be seen, the accuracy of the system is enhanced as the number of different used videos is increased. Hence, it is more important to use a diverse set of data, for example using more procedures, than using a large number of images from the same procedure.

Finally, Figure 7.5 shows a set of anchor images acquired with PillCam SB3, the target domain, and its top more similar images from the PillCam SB2 dataset, the source domain.

Table 7.4: Accuracy of the proposed system evaluated on *SB3D* trained with 600 from *SB3D* using different number of procedures.

Method	**SB3 Videos**	**Accuracy** (%)
	1	80.2
	2	85.7
TL_SB2-3	3	86.8
	4	86.8
	5	86.9

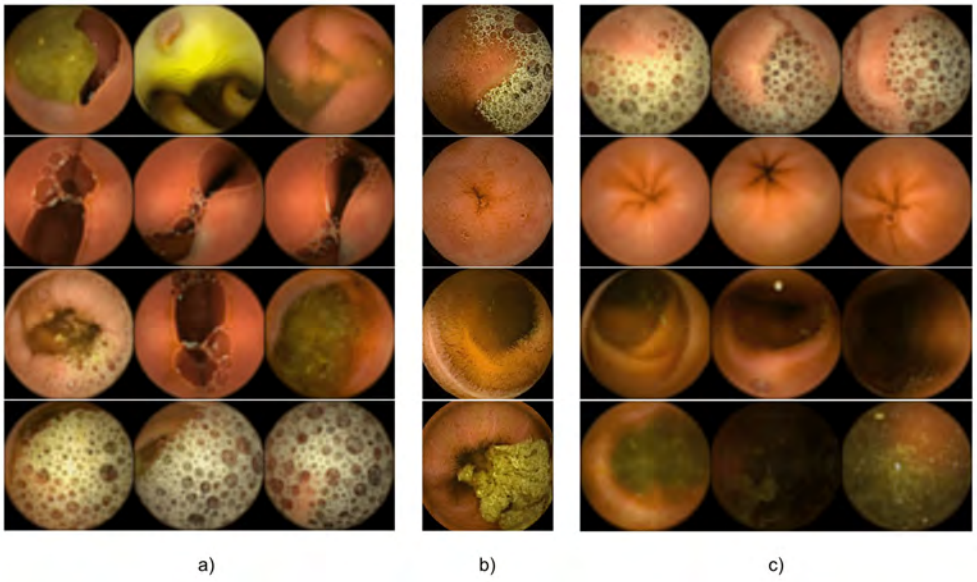


Figure 7.5: Each row shows a query where (b) is the anchor images from the target domain *SB3D* capsule, (a) the three most similar images to the anchor image without adapting the model to the target domain, and (c) The three most similar images adapting the model to the target domain.

Central images in each row represent anchor images while the three images at the left are the top most similar images before adapting the domain, and the three images at the right are the top most similar images when the embedding has been adapted. As can be seen, similar images when using the adapted embedding are really similar in shape and color to the anchor images, although their look and feel are blurrier.

7.6 Conclusions

In this work, we have explored the use of deep metric learning, based on the triplet loss function, to improve the generalization of a model over different datasets from different versions of WCE capsules. The proposed method is trained using a larger dataset from a

source domain, using an old WCE device, and is adapted to work on a target domain that represents images obtained by a new WCE device, with minimal labeling efforts. Results show evidence that with just a few labeled images from a newer camera, a model that has been trained with images from older systems can be readily used in the new environment.

We also explored, evaluated, and compared several different transfer learning solutions when dealing with small target domain datasets. We have shown that the triplet loss function may be well suited for dealing with the problem of data distribution shift over different domains. Particularly, we study the effects of using different amounts of images and procedures, concluding that diversity is more important than the amount.

Acknowledgements

We want to thank Carolina Malagelada and Fernando Azpiroz from Hospital General de la Vall d'Hebron for their clinical insights, the team from CorporateHealth International ApS for their feedback and economic support and NVIDIA for their GPU donations. This work has been also supported by MINECO Grant RTI2018-095232-B-C21 and SGR 1742.

Chapter 8

Paper V

Self-supervised out-of-distribution detection in Wireless Capsule Endoscopy images

Arnau Quindós[a], Pablo Laiz[a], Jordi Vitrià[a] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

In Artificial Intelligence In Medicine, Elsevier, Volume 143, 2023. ISSN 0933-3657

doi: 10.1016/j.compbiomed.2023.102606

Impact Factor Journal: 7.011
Q1 Computer Science, Artificial Intelligence (32/145)

Contents

Motivation and Context

The accurate detection of GI lesions is crucial for effective diagnosis and treatment. However, it is not sufficient for a system to perform well on the task it was designed for; it must also be able to detect OOD images, which may indicate other severe conditions. This study focused on the development of a reliable OOD detector, which is necessary for the adoption of CAD systems in clinical practice.

The nature of WCE images is wide and heterogeneous, making it challenging for the system to learn what is normal. In addition, some endoscopic images are considered abnormal because of an abnormality in a small area of the image, even though the rest of the image is completely normal. In these cases, an OOD detector is likely to classify the image as in-distribution, because the anomaly cannot outweigh the in-distribution features of the image. Thus, the goal of this study is to develop an OOD detector that can identify large and small anomalies in images.

To address this challenge, the following study presents a novel patch-based self-supervised approach for WCE images. It does not require labels to detect OOD samples and consists of three stages. First, the system uses SSL to learn the embedding representations of WCE

image patches. Then, the patches are clustered using the K-means algorithm and the embeddings learned. Then, the clusters are used as pseudo-labels to train a patch classifier, which is employed along ODIN to detect the patches that are OOD.

The results of the experiments provide compelling evidence that the suggested approach is capable of detecting a wide variety of pathologies. This is a significant finding because the ability to accurately and quickly identify pathologies is critical for the proper diagnosis and treatment of diseases.

In addition, the comparison of the proposed method with other state-of-the-art techniques demonstrated its superiority in nearly all pathologies. This suggests that the proposed approach can be used as an effective alternative or complement to existing methods in detecting OOD samples. The superior performance of the presented method can be attributed to its ability to identify small patterns in the image data that are not readily apparent to other techniques.

Self-supervised out-of-distribution detection in Wireless Capsule Endoscopy images

Arnau Quindós[a], Pablo Laiz[a], Jordi Vitrià[a] and Santi Seguí[a]

a) Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain

8.1 Abstract

While deep learning has displayed excellent performance in a broad spectrum of application areas, neural networks still struggle to recognize what they have not seen, i.e., Out-of-Distribution (OOD) inputs. In the medical field, building robust models that are able to detect OOD images is highly critical, as these rare images could show diseases or anomalies that should be detected. In this study, we use Wireless Capsule Endoscopy (WCE) images to present a novel patch-based self-supervised approach comprising three stages. First, we train a triplet network to learn vector representations of WCE image patches. Second, we cluster the patch embeddings to group patches in terms of visual similarity. Third, we use the cluster assignments as pseudolabels to train a patch classifier and use the Out-of-Distribution Detector for Neural Networks (ODIN) for OOD detection. The system has been tested on the Kvasir-capsule, a publicly released WCE dataset. Empirical results show an OOD detection improvement compared to baseline methods. Our method can detect unseen pathologies and anomalies such as lymphangiectasia, foreign bodies and blood with $AUROC > 0.6$. This work presents an effective solution for OOD detection models without needing labeled images.

Key words: out-of-distribution; anomaly detection; deep learning; capsule endoscopy.

The paper has been re-typeset to match the thesis style.

8.2 Introduction

Wireless Capsule Endoscopy (WCE) is an endoscopy technique that is an alternative to the standard procedure originally developed by Iddan et al. (2000). This method presents a variety of advantages versus standard endoscopy due to being far less invasive, not requiring sedation, and risking fewer potential complications. WCE makes use of a small pill-shaped capsule with a camera inside, rather than the traditional long, thin, flexible tube with a camera at one of its ends. This capsule can be easily swallowed, upon which the camera records hours of intestinal video that a medical team can later view to diagnose any gastrointestinal condition.

Nevertheless, WCE videos can contain thousands of images per patient and must be screened by medical specialists. This is a time-consuming and complex process. Its repetitive nature might lead to missing pathologies or other important elements (Koulaouzidis et al., 2021). For this reason, artificial intelligence offers a clear opportunity to support this task (Yang, 2020; Robertson et al., 2021; Koulaouzidis et al., 2022).

The application of AI techniques has been thoroughly investigated for the detection of abnormal or suspicious images in WCE. Several works have been presented for the identification or segmentation of pathological conditions such as bleeding (Hajabdollahi et al., 2018; Aoki et al., 2020; Saraiva et al., 2021), polyps or tumors (Yuan et al., 2017a; Laiz et al., 2020; Yang et al., 2020; Saito et al., 2020; Falin et al., 2022; Gilabert et al., 2022), angiectasia (Leenhardt et al., 2019), ulcers (Aoki et al., 2019; Klang et al., 2020; Ribeiro et al., 2022), motility disorders (Malagelada et al., 2008), as well as methods for multi-pathology detection (Seguí et al., 2016; Ding et al., 2019; Guo and Yuan, 2020; Vieira et al., 2021; Adewole et al., 2021; Pascual et al., 2022a; Jain et al., 2021). Deep learning currently represents the state-of-the-art for most of these problems and has demonstrated promising results. Nevertheless, independent of the performance on the task for which these models were designed, the ability to detect unseen Out-of-Distribution (OOD) images is crucial, as such OOD images may correspond to other severe conditions. For example, a polyp is an abnormal growth of tissue that can evolve into cancer, and therefore, its detection can be highly beneficial. However, a system that accurately detects polyps but fails to identify advanced-stage tumors would not be desirable. Therefore, the development of reliable OOD detectors in addition to supervised detectors is necessary for adoption in clinical practice.

The nature of capsule endoscopy images is wide and heterogeneous, which challenges deep learning models to learn what is normal or in-distribution. Furthermore, some images are considered abnormal due to an anomaly in a small area of the image, despite the remaining image being completely normal. In these cases, an OOD detector will most likely classify those as in-distribution, as the anomaly cannot outweigh the in-distribution features of the image. Therefore, one of the goals of this work is to develop a detector that is able to identify small anomalies.

In this study, we introduce a self-supervised method derived from ODIN (Liang et al.,

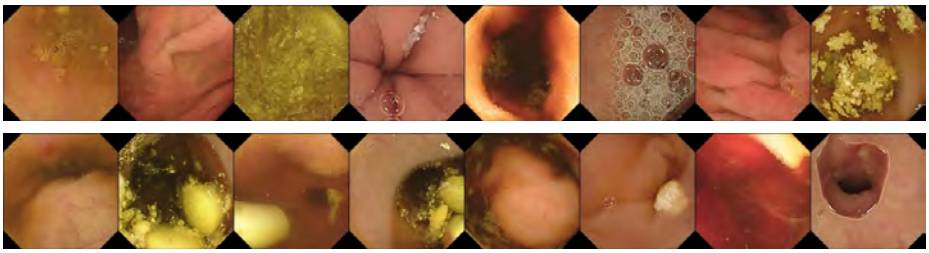


Figure 8.1: Random WCE sample images that illustrate the diversity of the dataset (Smedsrud et al., 2021) and the complexity of out-of-distribution detection. First row: normal frames. Second row: frames containing some pathology.

2017) based on patches. We first create a model able to generate vector representations of fixed-size patches extracted from WCE images as a self-supervised task. These embeddings encode visual features from the patches and allow the creation of clusters of patches in terms of visual similarity. Finally, we train a classifier using cluster assignments as pseudolabels. Similar to its predecessor method ODIN, our OOD detector is based on the confidence of this patch classifier.

The remainder of this paper is structured as follows. First, we present an overview of the related work in the OOD field. Then, we describe the details of our methodology, followed by the experimental setup and results. Finally, we conclude the paper and provide directions for future work.

8.3 Related Work

The OOD image detection problem in deep learning has been studied for many years using a variety of approaches ranging from conventional statistical techniques (such as density estimation) to generative models (such as autoencoders or GANs). In this study, we distinguish between supervised methods, which use some type of labeling in the training set, and self-supervised methods, which learn the necessary knowledge to perform the OOD problem without the need for specific labeling in the training set. Our method falls in the second category.

8.3.1 Supervised methods

A widely used baseline method for this problem is the maximum over softmax probabilities (MSP) (Hendrycks and Gimpel, 2016). This approach is based on a classifier trained over in-distribution data and works on the assumption that models make more confident predictions with in-distribution inputs than with OOD data. It conceptually depends on the outputs of a simple multiclass classifier and requires no further training. However, its performance has proven to be inferior to many later approaches. Thus, it is currently only used as a

baseline method.

Since diverse and enormous datasets of images are available, Hendrycks et al. (2018) proposed leveraging these data to improve OOD detectors against auxiliary datasets of outliers in a method called Outlier Exposure (OE). In this method, the classifier is trained to predict a uniform distribution over labels for outlier inputs, which enforces low confidence over these inputs. Thulasidasan et al. (2019) proposed using an abstention class in the classification problem and assigning known outliers to this class. Further work showed that leveraging the labels of the known outliers instead of assigning all outliers to a single abstention class can further enhance the performance of the OOD detector, despite only representing a small subset of the type of outliers that we want to detect Roy et al. (2021).

Another effective improvement to MSP is ODIN Liang et al. (2017). While still being a confidence-based approach, ODIN includes two fundamental novel techniques: temperature scaling and input perturbation. These techniques lead to better OOD detection, making it one of the best-performing state-of-the-art methods for the OOD problem. Nonetheless, ODIN relies on OOD data to tune the temperature and perturbation hyperparameters. In contrast, a generalized version of ODIN (GODIN (Hsu et al., 2020)) does not require tuning with OOD data and mitigates this issue.

Other approaches focus on modeling the class-conditional distribution of pretrained CNN features with a Gaussian distribution and use the Mahalanobis distance in the predicted class distribution to detect OOD samples (Lee et al., 2018). For example, DeepIF method (Li et al., 2020) achieves better detection performance by modeling the distribution of CNN features with a nonparametric technique based on isolation forests.

8.3.2 Self-supervised methods

The concept of learning normality to then detect anomalies is evident in methods based on deep-generative models including autoencoders (AEs), variational autoencoders (VAEs) and generative adversarial networks (GANs). All of these methods learn features with high representation quality that can be used for density estimation methods (Abati et al., 2018) or reconstruction error methods (Lu and Xu, 2018). These approaches rely on the assumption that reconstruction models trained on in-distribution images produce higher-quality outcomes with in-distribution inputs than with OOD inputs. Thus, images producing a high reconstruction error can be classified as OOD.

Other self-supervised approaches have tried to replicate classifier-based supervised methods without using labeled data, such as ensemble leave-out classifiers proposed by Vyas et al. (2018). This technique consists of randomly partitioning data in K subsets and creating K classifiers, each of which samples one of the K subsets without replacement as OOD data and samples the remaining subsets as in-distribution training data.

8.4 Methodology

The general concept of the proposed method in this paper is to use ODIN (Liang et al., 2017), which is considered one of the state-of-the-art approaches in OOD detection, to detect abnormal areas of WCE images. To focus on small regions of the image, we split the WCE images into fixed-size patches, which we consider our training and testing examples. Since ODIN is a classifier-based approach, labels are required to classify samples. Toward this application, we use a self-supervised feature extraction network to generate embeddings and then apply a clustering algorithm that assigns each sample one label that is later used to train the classifier. Importantly, our method does use any external labeling of the images or patches.

The pipeline of the method comprises three stages, which are henceforth described in detail.

8.4.1 Triplet-loss embeddings

The first stage of our method seeks to learn a vector representation for patches extracted from WCE images. To learn these embeddings, we use a Triplet Loss (TL) network, which allows us to perform self-supervised learning, as described in the next paragraph.

A TL architecture compares an anchor input with two other inputs: a positive input, which shares a property with the anchor, and a negative input, which does not share this property. In our case, inputs are fixed-size patches extracted from WCE images, and the shared property is that both subsets are extracted from the same image, whereas the negative patch is extracted from a different image, as illustrated in Figure 8.2. TL aims at ensuring that the anchor image, x_a, is closer to all other images from the same class, x_p, than any image from a different class, x_n.

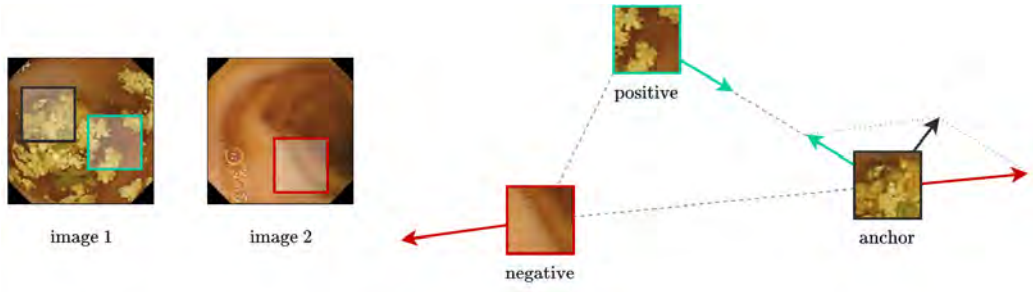


Figure 8.2: Triplet loss applied to three patches that are transformed into three vectors. Anchor: patch from a given image. Positive: a different patch of the same image. Negative: a patch of a different image.

To achieve classification, the following loss function is used:

$$\mathcal{L}_{TL}(x_a, x_p, x_n) = \sum_{i=1}^{N} \max\left(||f_i^a - f_i^p|| - ||f_i^a - f_i^n|| + \alpha, 0\right) \tag{8.1}$$

where f^k is the vector representation of x^k, N is the batch size, $||\cdot||$ is a norm and α is a margin parameter to enforce separation between classes.

Given two patches from the same image, this network generates embeddings that are closer together than two patches from different images. Since two patches from the same image will tend to be more visually similar than two from different images, these embeddings can be useful to cluster patches based on visual features.

8.4.2 Cluster pseudolabeling

As outlined above, we must label patches to train a patch classifier. Therefore, we use cluster predictions as pseudolabels to train our OOD-detector classifier.

Given the triplet-loss embeddings produced in the previous stage, we use the mini-batch K-means algorithm to create patch clusters based on visual similarity. Images in the same cluster tend to share visual features such as color, texture or shape. An example of such a clustering is shown in Figure 8.3.

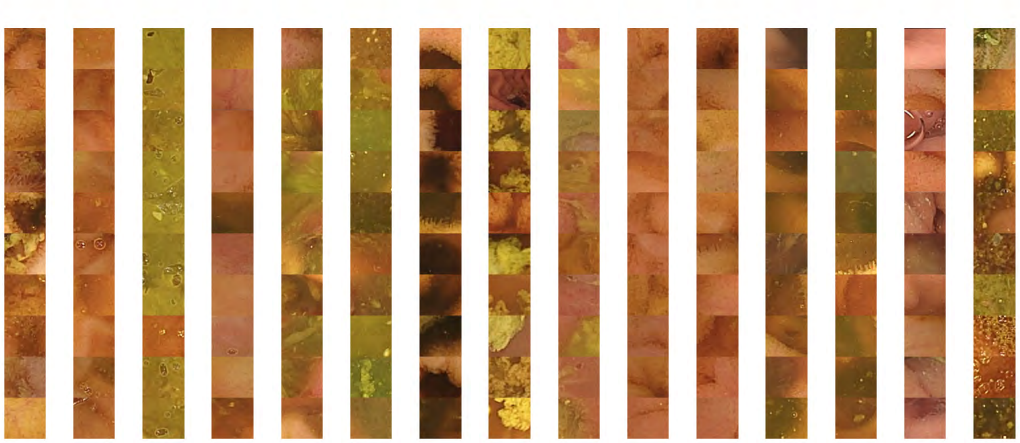
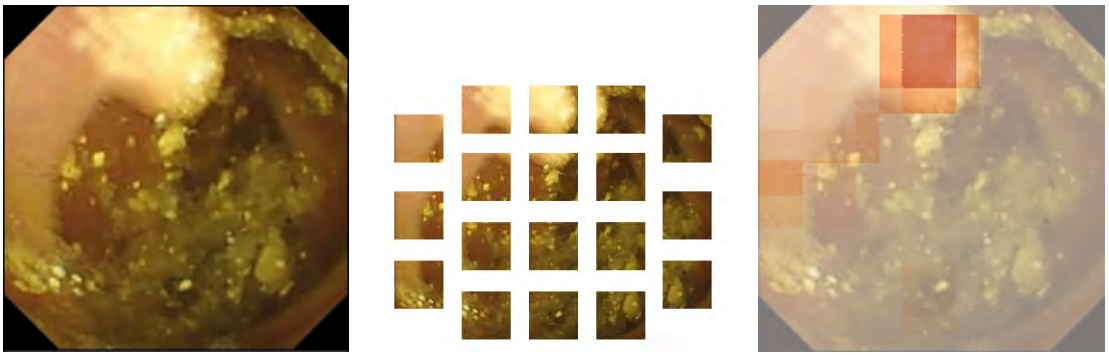


Figure 8.3: Example of clusters produced; each column represents one cluster. Patches in the same cluster are more visually similar than patches in different clusters ($K = 15$, patch size 96×96).

This clustering partitions the patch dataset, which is then used to train a K-class classifier.

8.4.3 Patch OOD classifier

The third and final stage of our method is the patch-based ODIN, which is based on a K-class classifier trained with the aforementioned pseudolabels. This classifier also includes temperature scaling and input perturbation, as defined in the original ODIN paper (Liang et al., 2017). A given image is partitioned into m patches and fed to the ODIN model, which outputs an anomaly score for each patch (see Figure 8.4).



(a) WCE example image with lymphangiectasia.

(b) Patches extracted from the image. Note some overlap between patches.

(c) Heatmap produced by ODIN. Red represents high anomaly score areas.

Figure 8.4: Illustration of the patch-splitting process.

Formally, we let x be an input patch, $\tilde{x}$ be the perturbed version of this patch and $z = (z_1, \ldots, z_K)$ be the output vector produced by the temperature-scaled K-softmax layer. We define the anomaly score, called the softmax score, in Equation 8.2. Then, these scores are combined using a summary function to obtain a measure of the abnormality of the image as a whole. If this score surpasses a certain threshold, then the image is labeled as OOD.

$$\mathcal{S}(\tilde{x}; T) = 1 - \max_{i=1,\ldots,k} softmax(\tilde{x}; T)_i = 1 - \max_{i=1,\ldots,k} z_i \tag{8.2}$$

For each image x, we will extract m patches $x_1, \ldots, x_m$. Given a perturbation magnitude ε, a temperature parameter T and a threshold δ, our OOD discriminator is defined as follows:

$$OOD(x; T, \varepsilon) = \begin{cases} 1, & \Psi\left(\mathcal{S}(\tilde{x}_1; T), \ldots, \mathcal{S}(\tilde{x}_m; T)\right) \geq \delta \\ 0, & \Psi\left(\mathcal{S}(\tilde{x}_1; T), \ldots, \mathcal{S}(\tilde{x}_m; T)\right) < \delta \end{cases} \tag{8.3}$$

where Ψ is a summary function applied to the softmax scores of the patches.

Given the softmax scores of the patches of an image $\vec{y} = (y_1, \ldots, y_m)$, we define three summary functions in Table 8.1.

Each of these three strategies may perform differently depending on the nature of the anomalies to detect. For instance, *max*, which only uses the patch with the highest anomaly

Table 8.1: Definition of the summary functions used in this paper. S_n is the subset of the first n softmax scores, sorted in descending order.

max maximum-score patch	$\Psi_{\max}(y) = \max_{i=1,\dots,m} y_i$
top-k average of top-k patches	$\Psi_{\text{top } k}(y) = \frac{1}{k} \sum_{y_i \in S_k 1} y_i$
wavg weighted average of all the patches	$\Psi_{\text{wavg}}(y) = \frac{1}{m} \sum_{y_i \in S_m 1} \lambda^i y_i$

score, might work better for localized features but might also introduce more noise; *wavg*, which accounts for all the patches but gives more importance to the highest scores, might be better suited for global anomalies; *top-k* is an intermediate approach that might work like *max* but with less noise.

8.5 Experimental setup

8.5.1 Dataset

We evaluate and compare our proposed method with Kvasir-Capsule (Smedsrud et al., 2021). Kvasir-Capsule is a publicly released WCE dataset that contains 117 videos of gastrointestinal footage from different patients, 74 of which are unlabeled and 43 partially labeled. The labeled frames are comprised of 14 different classes, 5 of which refer to non-pathological categories: Normal Clean Mucosa (NOR), Ileocecal Valve (IV), Pylorus (PYL), Reduced Mucosal View (RED) and Ampulla of Vater (AV); and 9 that refer to pathological or abnormal categories: Angiectasia (ANG), Erythema (ERY), Blood - fresh (BLO), Blood - hematin (BLH)*, Erosion (ERO), Foreign body (FB), Ulcer (ULC), Polyp (POL) and Lymphangiectasia (LYM). For our OOD problem, we considered the 9 pathological categories as OOD, i.e., our detection target.

Different frames from the same video can be very similar. Thus, data partitions must be done by videos instead of frames. We randomly selected 64 out of the 74 unlabeled videos for the training of the triplet network, the K-Means clustering and the patch classifier. The 10 remaining unseen videos are used as an intermediate validation set to assess the quality of the resulting embeddings and clustering and the accuracy of the classifier. The 43 labeled videos are then used only for testing purposes, with normal classes considered in-distribution and pathological frames of OOD. We extract fixed-size patches of 96×96 pixels from a video frame resolution of 336×336 using a step size of 60 pixels between patches, while ensuring that there is overlap between patches and that all areas of the image are captured.

*BLH class is not considered for evaluation purposes due to the small number of frames available in this category, of which there are only 10.

8.5.2 Method stages

1. **Triplet-loss embeddings** The triplet-loss model uses the EfficientNetB0 (Tan and Le, 2019) architecture, followed by a global average pooling layer. Finally, a 1280-unit dense layer outputs the feature vectors.

2. **K-Means clustering** Due to the high volume of images, we used the mini-batch version of the K-Means algorithm. The value chosen for the number of clusters is $K = 20$ due to seemingly showing the most consistent clustering results.

3. **Patch ODIN classifier** We use a CNN classifier that comprises the first three blocks of the ResNet50v2 (He et al., 2016) architecture (ImageNet pretrained), followed by three fully connected layers integrating dropout and batch normalization layers. The top fully connected layer uses a temperature scaling with a fixed temperature parameter value $T = 1000$, as proposed in other previous work (Hsu et al., 2020). Because the number of labeled videos for each pathology is limited, cross-validating these parameters is very risky.

8.5.3 Baseline methods

We use baseline self-supervised methods trained with the same data to compare the performance of our approach with other previous work. The implementation of supervised approaches would require using additional data or labels and, more importantly, would make an unfair comparison. The following methods are used as baselines:

ODIN (Liang et al., 2017). To incorporate a nonpatch-based ODIN as a baseline method, we trained a self-supervised model that uses temporal information of the frames to generate pseudolabels and train an ODIN classifier on the full image. Importantly, this adapted version is self-supervised, indicating that it has not been trained with ground-truth labels.

SelectiveNet (Geifman and El-Yaniv, 2019). The SelectiveNet architecture includes a rejection option for selective classification, which we use as the OOD score. We use the same self-supervised methodology to generate pseudolabels as described for ODIN.

VAE (Lu and Xu, 2018). We use the same VAE architecture and train on the same in-distribution data as our approach. Then, we use the VAE anomaly score proposed in their work to determine if an image is an OOD.

Patch VAE. We use a patch-based method, but instead of ODIN, we introduce the VAE from the baseline method (trained on patches) to create an anomaly score for every patch. Then, for every image, the scores of the patches are combined using a summary function to obtain an anomaly score for the image. We use this model as an intermediate between the VAE and our method.

8.5.4 Evaluation metrics

- **Accuracy.** Measures the fraction of examples correctly classified.
- **True Negative Rate (TNR)** or **specificity**. Refers to the proportion of negative examples that were classified as negative: $TN/(TN + FP)$.
- **True Positive Rate (TPR)** or **sensitivity**. Refers to the fraction of positive examples that were classified as positive: $TP/(TP + FN)$.
- **AUROC**. Area under the ROC curve. Refers to the sensitivity-specificity tradeoff at various threshold settings. To determine AUROC, we compute the anomaly score of normal and OOD samples and measure sensitivity and specificity at TPR and FPR at different threshold configurations.
- **TPR at n% TNR**, abbr. TPRn. Refers to the TPR when the TNR is $n\%$. TPR95 and TPR90 are used.
- **%PF**. Refers to the percentage of pathological frames among all the frames in a video.
- **%DPF(n)**. Refers to the percentage of pathology among the n frames with the highest outlier scores.
- **Difference between %PF and %DPF (*diff*)**. Given the high variance in %PF across different videos in the dataset, we use this to measure how well the model detects OOD frames for different pathological prevalences.

$$diff = \%\text{PF} - \%\text{DPF}$$

In addition to these quantitative metrics, we also evaluate the system qualitatively by inspecting the results produced on a subset of images. We consider this a very important evaluation to understand the predictions obtained by each model, which allows us to understand which images the model considers abnormal and which others are classified confidently.

8.6 Results

In our first experiment, we seek to analyze anomaly score distributions produced by the patch ODIN [†] for each class. Toward this goal, we extract the softmax score of each image and fit a Gaussian distribution to each set of scores. Figure 8.5 shows these normalized distributions, i.e., with balanced classes to better compare the degree of overlap of these distributions. We note that, as a result of the plain nature of ERY images in this dataset, this class produces even lower anomaly scores than the normal class. The normal class

[†]For the patch ODIN, we fix the following parameters: $k = 20$, $\varepsilon = 5 \cdot 10^{-4}$, and $T = 1000$.

(NOR) produces the lowest anomaly scores and thus allows us to separate classes using these scores. However, the degree of separation varies for each pathology: some classes, such as ERO and BLO, have a large overlap with NOR, while others, such as LYM or ULC, have a significant separation. We compare each pathology versus the normal class to evaluate the potential of our OOD discriminator. The results for a patch size of 96 are shown using ROC curves in Figure 8.6.

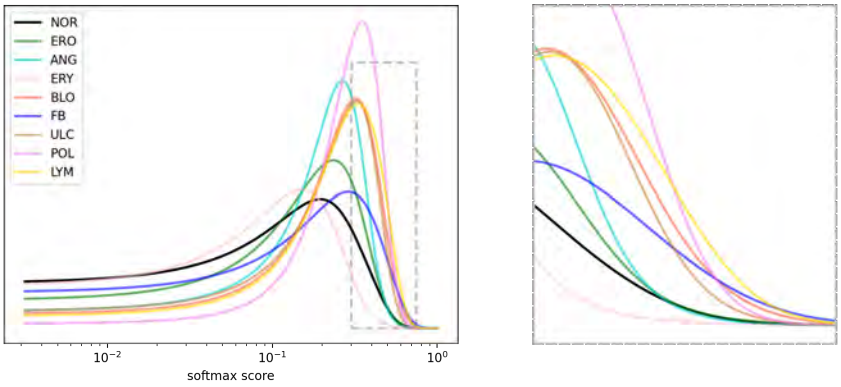


Figure 8.5: Softmax score distributions produced by patch ODIN (*top5* summary) by pathology. Real distributions are used to fit Gaussian distributions, which are plotted above. The second window shows a zoomed view. Distributions are normalized for the sake of comparison.

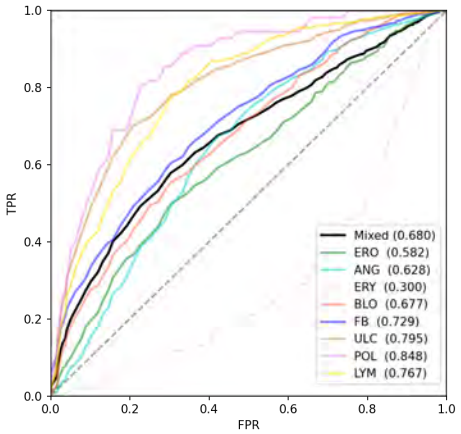


Figure 8.6: ROC curve of patch ODIN (*top 5*) OOD detection by pathology. Mixed considers all Kvasir pathological frames as one single class.

To further investigate the effect of the patch size, we repeated the experiments using additional patch sizes of 64 and 128. The results are presented in Table 8.2, which shows that the patch size of 96 yields the best results. Therefore, we adopt this size for the subsequent experiments.

The AUROC score by pathology and comparison with baseline methods are summarized in Table 8.3. The results show that, considering all pathological frames as one abnormal

Table 8.2: AUROC scores of OOD detection by pathology of the proposed Patch ODIN method. Comparison between three different patch sizes (PS).

pathology	**#samples**	**PS** 64×64			**PS** 96×96			**PS** 128×128		
		max	*top-k*	*wavg*	*max*	*top-k*	*wavg*	*max*	*top-k*	*wavg*
ANG	866	0.576	0.629	0.636	0.572	0.628	0.635	0.592	0.591	0.595
BLO	446	0.646	0.708	0.701	0.613	0.677	0.678	0.644	0.616	0.601
ERO	507	0.579	0.593	0.598	0.562	0.582	0.587	0.538	0.557	0.564
ERY	159	0.412	0.398	0.407	0.322	0.301	0.300	0.457	0.468	0.479
FB	776	0.669	0.698	0.696	0.737	0.729	0.732	0.621	0.613	0.602
LYM	592	0.645	0.671	0.671	0.739	0.767	0.772	0.627	0.635	0.640
POL	55	0.531	0.417	0.435	0.776	0.848	0.845	0.580	0.600	0.600
ULC	854	0.615	0.618	0.626	0.725	0.795	0.804	0.534	0.521	0.518
Aggregated	4255	0.611	0.638	0.640	0.650	0.680	0.686	0.578	0.578	0.575

class, our method slightly improves performance over the baseline methods. Considering each pathology individually, we observe different results. Patch ODIN performs especially well with LYM and ULC and slightly outperforms the baseline methods with FB and ERO. However, our method does not improve BLO detection, which the VAE model does especially well. This is attributed to blood being the most global anomaly, such that splitting the data into patches does not contribute to better detection. Furthermore, we observed that the best summary strategy depends on pathology. Some of the pathologies are global, while others appear very localized; overall, *top5* and *wavg* seem to yield the best results.

Table 8.3: AUROC scores of OOD detection by pathology, comparison between different methods. For each pathology, the best score is marked in bold.

pathology	**#samples**	**Patch ODIN** *96×96, top k*	**Patch VAE** *96×96, top k*	**ODIN**	**VAE**	**SelectiveNet**
ANG	866	**0.628**	0.367	0.483	0.573	0.515
BLO	446	0.677	0.705	0.541	**0.791**	0.576
ERO	507	0.582	**0.622**	0.570	0.540	0.472
ERY	159	0.301	0.231	**0.560**	0.324	0.326
FB	776	**0.729**	0.623	0.642	0.679	0.632
LYM	592	**0.767**	0.671	0.752	0.745	0.738
POL	55	**0.848**	0.350	0.667	0.622	0.652
ULC	854	**0.795**	0.706	0.543	0.680	0.669
Aggregated	4255	**0.680**	0.577	0.578	0.642	0.572

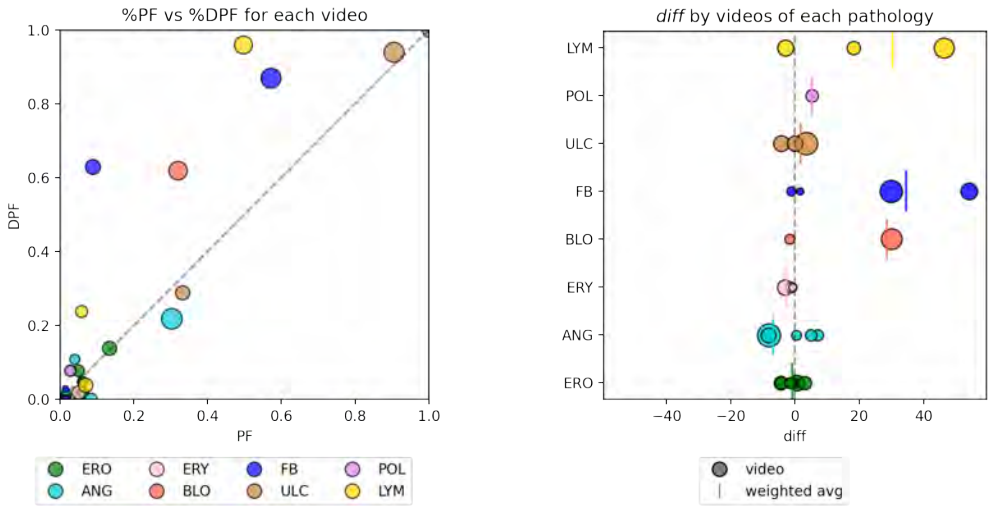
In general, we observe that VAEs tend to assign higher anomaly scores to images that appear more complex in terms of texture, colors and shapes. For instance, we find that nonpathological bubble images are usually assigned high scores, while pathological plain images are not detected. This mainly occurs because complex features, despite being common in the training set, are harder to reconstruct for an autoencoder. Thus, reconstruction

error is higher for complex images, which plays a large role in anomaly score.

The availability of images for certain pathologies is extremely limited (different images may be consecutive frames of the video that contain the same anomaly), which can lead to inaccurate results. For this reason, further qualitative and quantitative analysis is necessary to confirm the performance of the system.

Notably, to compute AUROC scores, we use normal and OOD frames extracted from different videos. In real-world situations, given a WCE video from a single patient, it is desirable to flag the most abnormal frames to detect any potential condition. To better measure the performance of the model in such a situation, we test our method in each video separately and measure how well the model detects pathological frames among the 'most abnormal' frames. For this test, we use the *diff* metric described in the previous section, with $n = 100$. This metric compares the percentage of pathological frames among the 100 frames with the highest outlier score (%DPF(100)) with respect to the percentage of pathological frames in the video (%PF).

The results of the video analysis are shown in Figure 8.7. We observe that frames containing LYM, FB and BLO produce high anomaly scores and thus are detected among the most outlier frames. For remaining diseases, the average *diff* is close to 0, indicating that the model detects pathological frames (among the most abnormal) at the same rate as they are present throughout the video. Because labeled videos containing ULC are ¿90% pathological, *diff* may not be the best performance indicator for this class. Additionally, these results may not match with the AUROC scores as previously presented because this metric measures each video independently and focuses on the most abnormal tail end.



(a) %PF vs. %DPF ($n = 100$)

(b) *diff* by pathology

Figure 8.7: Results by video. Each circle represents a video, and the size is proportional to the number of frames contained in that video.

We also conducted a qualitative analysis using the outputs of the model on a subset of images. To do this, for each selected image, we examined the score of each patch, analyzed which patches produced the highest scores, and plotted the results in the form of a heatmap over the original image. This process is illustrated in Figure 8.8, where the model performs well, and in Figure 8.9, where the model fails to correctly identify anomalies. We examined both successful and unsuccessful examples to determine those types of anomalies our model is able to identify and those which it cannot. A general conclusion is that the model tends to detect more visually prominent anomalies more accurately, as was expected.

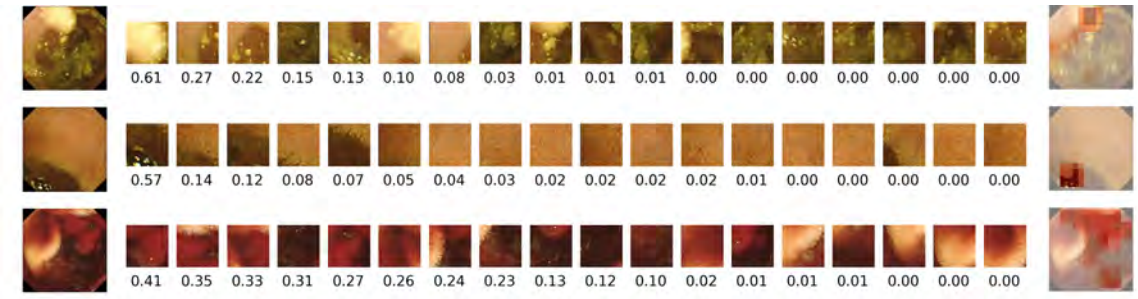


Figure 8.8: Left: sample WCE images that contain LYM, FB and BLO, from top to bottom. Center: patches extracted from each image sorted in terms of softmax score. Right: Heatmaps produced using softmax scores; red areas represent high anomaly scores. In these examples, the model correctly identifies anomalies, and thus, patches containing anomalies produce high scores.

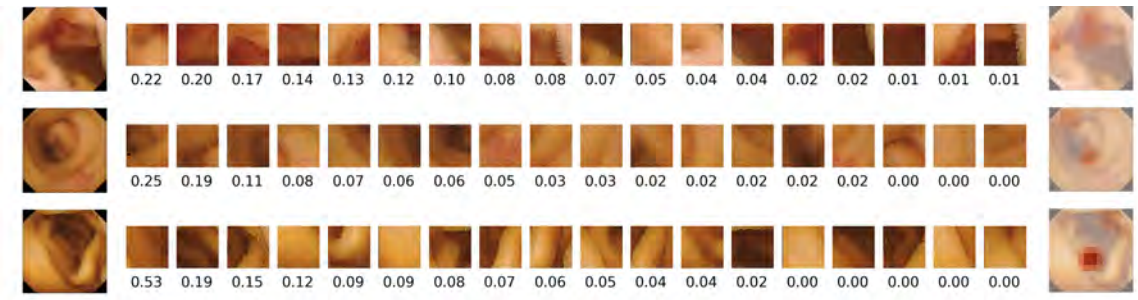


Figure 8.9: Left: sample WCE images that contain BLO, ANG, and no pathologies (NOR), from top to bottom. Center: patches extracted from each image sorted in terms of softmax score. Right: Heatmaps produced using softmax scores; red areas represent high anomaly scores. In these examples, the model is not able to correctly identify anomalies. In the first two cases, all the patches are assigned low scores, and thus, any abnormal area is detected. In the third case, the model incorrectly assigns a high score to a normal patch.

8.7 Conclusions

This study presents a method to improve OOD detection in WCE images with respect to other self-supervised approaches, such as VAEs or ODIN without patches. Both quantitative and qualitative results show that the system successfully detects pathologies including lymphangiectasia, foreign bodies, and blood showing. Moreover, the patch-based nature of our methodology allows us to measure the abnormality of every region of the image.

While our method tends to effectively detect the most visually prominent anomalies, it is less sensitive to subtler anomalies such as erosion, angiectasia or erythema. These pathologies are visually quite similar to in-distribution WCE images. Therefore, detecting these types of anomalies using a model that has not been previously exposed to them is a great challenge. Moreover, the limited availability of data in medical fields reinforces the need for larger and more diverse datasets. Toward this goal, our future work may consider incorporating online learning techniques, where the model could dynamically adapt as a medical team flags unseen images to the system.

Overall, we intend that our method can provide an effective solution for OOD detection models without the need for labeled images. While this work has focused on WCE images, our methodology can also be applied to OOD detection in other computer vision applications.

Acknowledgments

This work has been partially funded by MINECO Grant RTI2018-095232-B-C21, SGR 01104 (Generalitat de Catalunya), Innovate UK project 104633, AI AWARD 02440 project from NIHR and NHSX, and the FPU grant FPU17/02727 (Ministerio de Universidades, Spain).

Chapter 9

Discussion of the Results

Contents

This chapter provides an analysis of the outcomes obtained in the different studies detailed in this thesis. The presented research comprises five individual papers, which can be categorized into three categories: pathology detection, anatomical landmark identification, and OOD sample handling.

9.1 Pathology Detection

The first addressed problem is the detection of pathologies, mainly focusing on colorectal polyps. The proposed CAD systems aim to identify them with high precision based on their visual attributes. Moreover, this problem is associated with three DL challenges: scarcity labeled data, imbalanced datasets, and model explainability.

9.1.1 Paper I

Commonly, the pathology detection problem only involves solving an image-classification task. However, the methodology explained in the first paper to detect polyps(Chapter 4) proposed another optimization strategy to achieve the goal and mitigate the small and imbalanced dataset issues. Specifically, the TL with the Batch All strategy for mining the triplets (Section 3.1.2) was used to train the backbone network and obtain informative embeddings of the images. Subsequently, a dense layer was employed over the embeddings to produce the predictions of the system. This layer acted as a classifier and was optimized with the CCE loss.

Although full details about the dataset used are presented in Chapter 4, the main ones are summarized in the following lines. 120 procedures were employed, where only 52 contained at least one polyp. The total number of unique polyps in the dataset was 165 distributed in $2,136$ frames.

Quantitative and qualitative experiments were conducted to analyze the performance of the proposed approach. First, several experiments were run to explore different hyperparameters of the system, including the margin m of the TL and the size of the low representation of the images. These tests helped to identify the best configuration for the task, concluding that the appropriate margin and embedding size were 0.2 and 2048, respectively.

After obtaining the best experimental setup, the model was compared against two baselines: ResNet, the backbone of the system trained only with CCE loss, and the proposal with Batch Hard strategy mining. The results showed that the proposed method outperformed both baselines, obtaining an AUC-ROC of $92.94 \pm 1.87\%$. In addition, the sensitivity scores for the specificity at 95%, 90%, and 80% were $76.68 \pm 4.93\%$, $82.86 \pm 4.78\%$, and $88.53 \pm 3.76\%$, respectively. These scores represent the percentage of polyp images detected by revising only the top 5%, 10%, and 20% of video frames.

In the final quantitative experiment, the designed method was compared against state-of-the-art models, Yuan and Meng (2017), Guo and Yuan (2019), and Yuan et al. (2020). The results showed that the proposed model was more efficient in the retrieval task. In particular, it outperformed each metric by at least 2%.

It is important to note that the commented experiments were frame-based. Consequently, the metrics measure whether the system detected each frame in which a polyp

was visible. However, for physicians, it is crucial to find all unique polyps, which is not the same as finding all the frames from each polyp. Hence, the results were also displayed and discussed considering the identification of unique polyps. Briefly, detection at different specificity values, 95%, 90%, and 80%, were above 90% regardless of their size or morphology.

Finally, the implementation of CAM (Zhou et al., 2016), an explainability method, in the system allowed the extraction of qualitative results. After visualizing several examples of true positives, false positives, and false negatives, several conclusions were drawn. True and false positives demonstrated that the system had learned the visual features of polyps. However, false negatives were difficult to analyze using only one frame, which is also complex even for expert gastroenterologists. Overall, this incorporation might help physicians locate polyps in the images.

In summary, the TL proved to be appropriate for addressing all the mentioned challenges in this problem. Moreover, the use of CAM might allow experts to know the location of the polyp and increases their trust in the proposed CAD system. Despite the good results obtained, there is still room for improvement. With the development of new DL methodologies, more powerful embeddings can be extracted. Moreover, as the Batch All mining strategy uses random images from each class, other techniques might be designed to provide large diversity between the samples in each batch. This improvement would enhance the sampling process and the generalization of the network.

9.1.2 Paper II

The second paper, included in Chapter 5, deals with the detection of polyps and the identification of inflammatory and vascular lesions. As described in Section 2.3.2, the lack of labeled data and the imbalanced dataset are usually mitigated using transfer learning, particularly with ImageNet weights. In this study, a SSL method (Section 3.2) was proposed to generate a proper initialization for the network weights. To this end, the method leveraged the temporal axis of WCE videos to obtain richer embeddings. Specifically, the developed SSL method learned to produce similar representations for images belonging to the same sequence and distinct embeddings for images belonging to different sequences.

To perform this task, 49 unlabeled WCE videos were used. The SSL was performed with the TL and sequences of 9 images. The qualitative results showed that the method successfully generated rich representations. In addition, a set of experiments were performed to optimize the hyperparameters of the system: sequence size, sequence batch, window size, number of projection layers, and projection dimensionality. The best setup obtained was with hyperparameters 72, 1, 9, 3 and 128, respectively.

Once the weights of the network converged to a solution of the pretext task, two downstream tasks were applied. The first one aimed to detect polyps employing the methodology and the dataset of the previous article (Laiz et al., 2020). The proposed method achieved

state-of-the-art results compared to the previous paper and the well-known SimCLR method (Chen et al., 2020a), demonstrating its ability to handle imbalanced datasets. The method reached an AUC-ROC of $95.00 \pm 2.09\%$ and sensitivity scores at specificity of 95%, 90%, and 80% of $80.16 \pm 6.97\%$, $86.31 \pm 6.20\%$, and $92.09 \pm 4.63\%$, respectively. The second task involved a 3-class classification problem. It consisted of a dataset formed by 600 images from normal, inflammatory, and vascular lesions; hence, $1,800$ images. The proposed method achieved comparable results to the current state-of-the-art method (Guo and Yuan, 2020) and proved that the approach can also deal with small datasets.

However, this system has two main limitations. First, sequences of images are required to train the SSL model, which limits its applicability in certain medical imaging domains, such as X-rays. Second, because of hardware limitations, the number of images that can be used during training is limited, leading to less diverse learning embeddings. To overcome these limitations, future research could explore alternative SSL architectures to improve the downstream task performance and use different hyperparameter settings and sampling techniques. In addition, expanding the system to other WCE domains and medical fields would be highly beneficial.

9.2 Anatomical Landmark Identification

The second problem is the identification of the anatomical landmarks of the small bowel and large intestine. This task entails distinguishing between structures in the digestive system with similar appearances.

9.2.1 Paper III

In Chapter 6, a clinical decision-support tool was proposed. The system performed two tasks, namely, image classification and anatomical landmark identification. The first aimed to predict whether the image was between the landmarks, while the second one identified the exact frame where the anatomical landmark was located.

The designed method employed a combination of still images and timestamp information to predict the probability that an image belonged to a specific organ. To improve the accuracy of these predictions, a bidirectional RNN was used. It smoothed the probability scores by incorporating context, time, and motion data. The output was utilized to measure the performance in the image classification task. Then, to identify the anatomical landmarks, a minimization problem was applied to the smoothed signal to determine the beginning and end of the organ.

The performance of the system was evaluated using three datasets, two from the small bowel and one from the colon. The first one was a public database, called *Kvasir-Capsule* (Smedsrud et al., 2021), containing 24 small bowel WCE studies. Each video had a mean of $44K$ frames on average. The second one, named *VH* and also extracted from small intestine capsules, was formed by 48 explorations with a mean number of frames of $35K$. Finally, 68 colon WCE studies formed the last dataset, *Capri*. The average of frames, in this case, was $14K$.

In the first set of experiments, the image classification task was evaluated. Initially, an ablation study was carried out to determine the effects of each component on the system. The results showed that adding the timestamp enhances the performance with respect to training only with still images. Moreover, considering the context, temporal, and motion information, the system further improved, achieving at least an AUC-ROC value of $96.00 \pm 4.57\%$ and a Mean Accuracy (MACC) score of $87.47 \pm 7.49\%$. The proposed method was also compared with state-of-the-art models: Zou et al. (2015), Chen et al. (2017a), Zhao et al. (2021) and Son et al. (2022). The results showed that the developed system outperformed previous methods in all datasets.

The second set of experiments focused on identifying the anatomical landmarks. This task was measured as the error between the predicted and actual frames. As in the previous experiments, the ablation study showed that incorporating context, temporal, and motion information with the probabilities of the images improved the performance of the method. When compared to the state-of-the-art methods Zhao et al. (2021) and Son et al. (2022),

the proposed method exhibited the lowest error. Moreover, a comparison of the landmark identification strategy was performed, revealing that the designed approach produced more accurate predictions.

The developed system is restricted to the analysis of one organ, owing to the lack of multi-organ labels in the datasets used. To overcome this limitation, future research should incorporate datasets with several labeled organs and adapt the CMT block accordingly. Furthermore, future methods should also accommodate different devices. In addition, to enhance the system, future research could explore the localization of other landmarks, such as the flexures of the large intestine. By incorporating these improvements, the proposed system could be better equipped to handle a wide range of medical scenarios and provide more accurate diagnoses.

9.3 Out-of-Distribution

The final problem faced in this thesis is related to OOD samples. As explained in Section 3.3, OOD includes all cases in which the data distributions from the training and test sets are different.

9.3.1 Paper IV

In Chapter 7, a novel method addressed the covariate distribution shift (Section 3.3.1) that arises from different capsule devices. This study aimed to adapt the domain of both device to a common space. To achieve this, the proposed method used the TL (Section 3.1.2) to ensure that embeddings from images belonging to the same class, but different devices, remained closer than any other embedding of an image from the same domain.

Two datasets were used in this study, one from PillCam SB2 (old device) and one from PillCam SB3 (new device). Both datasets contained images that were classified into six classes. However, the first dataset consisted of $20,000$ images per class extracted from 50 procedures, while the second had only $1,000$ images per class captured from 10 procedures.

The first set of experiments was aimed at detecting the classification performance of the system in different scenarios. The results showed that the model trained on the old dataset and evaluated on the new one dropped drastically from an accuracy of 92.50% to 51.70%. Transfer learning was then applied, which resulted in a considerably good performance on the new dataset, with an accuracy of 87.00%, but a worse performance on the old dataset, 62.70%. Finally, the proposed domain adaptation method achieved the best results for the new dataset, 89.30%, and enhanced the results for the old dataset to 93.10%.

After proving the proposed approach as a good candidate for addressing the targeted problem, the impact of using different numbers of images from the new dataset and the number of procedures was explored. The results showed that increasing the number of images employed in the training improves the obtained accuracy. In addition, experiments suggested that the diversity of images (number of different procedures) was more important than the number of images. These findings were supported by the qualitative results. Moreover, it was observed that the system learned about lesion features, but not device features, when using the TL for domain adaptation.

A limitation of the proposed method is the need for labeled data to perform domain adaptation. Future work should focus on using unsupervised learning to avoid the requirement of labeled target data in the training process.

9.3.2 Paper V

Paper in Chapter 8 focused on the problem of detecting OOD samples. For this purpose, a three-stage SSL patch-based method was proposed. In the first step, a network was trained to extract the embedding of patches using a SSL approach (Section 3.2) with the TL (Section 3.1.2). Then, using the embeddings, a clustering method was applied to create pseudo-labels for a classification problem. Finally, a network was trained to classify the patches with the pseudo-label clusters. When this network performed properly, its outputs were used along with ODIN to detect OOD patches.

To evaluate this approach, the public dataset *Kvasir-Capsule* (Smedsrud et al., 2021) was used. The 74 unlabeled videos were used to train the classifier in a self-supervised manner. In addition, 43 videos, partially annotated in 9 pathological categories, were used to validate the proposed OOD detector.

The first experiment aimed to determine the optimal patch size. The results showed that patches of 96×96 pixels obtained the best performance among the proposed metrics, with an aggregated score of 0.68 with the *top-k* metric. Then, the method was compared with state-of-the-art approaches Lu and Xu (2018), Geifman and El-Yaniv (2019), and themselves with patches. The aggregated score obtained by the proposed method outperformed all systems. However, specific classes exhibited higher scores in other tested models. This happened with pathologies considered global lesions; hence, in this case, splitting the image in patches does not improve the detection. The same conclusions were obtained after analyzing the qualitative results.

To overcome the fact that general lesions may not be detected as OOD samples, other approaches for obtaining the OOD score have to be proposed and thoroughly explored. Additionally, other computer vision applications could benefit from the patch-based methodology utilized in this study.

9.4 Summary

Throughout this dissertation, five CAD systems have been presented, introducing innovative techniques to enhance diagnosis in WCE studies. The proposed models have addressed the challenges and limitations of both, DL and WCE, obtaining state-of-the-art results in each of their tasks. Therefore, the developed CAD systems have enhanced the effectiveness of WCE diagnosis.

In particular, the TL has allowed models to obtain richer embeddings and generalize better than with other standard approaches, even when data have been limited and imbalanced. This has been particularly beneficial in the pathology detection task.

The employment SSL has proven to be valuable as a first-step task to obtain more informative features by leveraging unlabeled data. The network trained with this methodology serves to be employed as a pretrained model, capable of transferring its knowledge to various other tasks. This approach is of great help when working with limited amounts of labeled data.

Finally, ODIN has played a crucial role in detecting pathologies that were not included during the training phase. Specifically, it has been demonstrated that the synergy between ODIN and SSL techniques allows the development of robust and reliable systems to detect unseen pathologies, and thus, improve the diagnosis of WCE videos.

Despite the promising results that CAD systems have shown, there are still many limitations and weaknesses that need to be addressed to further improve their performance. By addressing these issues, the research community will continue improving the reliability and capability of CAD systems for clinical use, which may lead to more accurate diagnoses and better patient outcomes. Therefore, DL applied to WCE studies will help to mitigate the current bottleneck of human pre-reading resources and enable medical professionals to diagnose and treat patients more efficiently and effectively.

Chapter 10

Conclusions

Contents

The content of this chapter is focused on two parts. First, the contributions of this thesis are summarized and compared with the goals proposed in Section 1.2. Then, future research directions are discussed.

10.1 Thesis Contributions

This thesis provides meaningful contributions to the development of state-of-the-art CAD systems. Specifically, they incorporate advanced DL techniques to improve diagnostic performance in WCE studies. These innovative methods have demonstrated remarkable results and might have the potential to advance the field of medical imaging. Therefore, it can be concluded that the main objective of this thesis has been successfully accomplished.

Moreover, the specific goals defined in Section 1.2 are also fulfilled.

1. **Create a model to tackle data scarcity**. By using the TL, SSL, and domain adaptation, the proposed DL methods have mitigated the problems caused by small datasets. This can be observed, first, in the polyp detection problem and then, in the domain adaptation proposal.

2. **Create a model to tackle imbalanced datasets**. Analogously to the previous goal, the TL and SSL mitigate the imbalanced dataset by enhancing the generalization of DL methods in the detection of polyps.

3. **Create a model that integrates explainability**. Through the use of CAM (Zhou et al., 2016), the CAD system for polyp detection provides the reason behind the prediction. Consequently, the trust of experts in the method could increase.

4. **Create a model to detect GI pathologies**. Two systems have been developed for detecting polyps. The first uses the TL to enhance the feature compression in the embedding. Whereas, the second method utilizes a SSL approach to obtain richer features from unlabeled WCE videos and then, applies this knowledge to the detection task. Finally, a third method has been implemented to detect several GI lesions without the need for labels.

5. **Create a model to identify anatomical landmarks**. Each of the required anatomical landmarks from the small bowel and the large intestine is detected in WCE videos by combining the image, timestamp, and motion information.

6. **Create a model that generalizes with data from different devices**. The covariate distribution shift has been addressed using a domain adaptation approach with the TL to align data from different capsule devices.

7. **Create a robust model to detect OOD samples**. An OOD detector has been implemented in a patch-based manner using SSL and ODIN.

Through extensive experiments and evaluations, it has been demonstrated the effectiveness of the proposed systems in the three problems. All papers have achieved state-of-the-art results and constitute a contribution to the medical field.

10.2 Future Lines of Research

The fields of medical imaging and WCE are rapidly evolving, and five approaches have been proposed to address their associated challenges. Although these methods and others from the literature may be successful, their limitations suggest that there is scope for further research. In this section, it is discussed some potential lines of work that are necessary to continue advancing in the field in the future.

For pathology detection, TL was shown to be valuable in mitigating small and imbalanced polyp datasets. Thus, it would be worthwhile to explore the same approach for the detection of other pathologies. Similarly, TL can also be applied to other medical imaging fields, where images with anomalies represent a small portion of the total dataset. It can also be used in multi-class problems to improve the feature-extraction process.

Most CAD systems currently use still frames to classify the image, whereas an expert uses the entire sequence to determine whether there is a pathology in the image. Therefore, it would be interesting to investigate the use of RNNs over a sequence that contains images with potential lesions. This could possibly improve the diagnostic accuracy of CAD systems.

Anatomical landmarks are crucial in the screening process, but current CAD systems only identify the start and end of an organ. To increase the quality of the video analysis, all landmarks from each organ should be identified. We believe that the proposed method, with proper data and a few adjustments, can identify these landmarks and improve the screening process.

WCE recorded data is readily accessible for further on-demand reviews. Therefore, these videos offer a unique learning platform for the new generation of physicians and researchers. This is an ideal scenario for the development and training of DNN systems and AI diagnoses. In the future, it would be worthwhile to continue researching the use of more powerful DL models to improve diagnostic accuracy and reduce the rate of missed lesions. In addition, the development of user-friendly software that integrates these models into clinical practice could facilitate the widespread adoption of these technologies (Leenhardt et al., 2019).

Finally, as the field of DL continues to evolve, it is likely that new and more advanced techniques will emerge to address the challenges faced in this thesis. For instance, recent research is been exploring the use of meta-learning, few-shot learning, and continual learning to improve the efficiency, scalability, and adaptability of diagnostic systems for healthcare. It is certainly an exciting time to be working in this field, as many more advances and breakthroughs are expected in the coming years.

Appendix A

Research Outcome

This thesis has led to the publications summarized below:

Journal Submissions

- **P. Laiz**, J. Vitrià, H. Wenzek, C. Malagelada, F. Azpiroz, and S. Seguí. WCE polyp detection with triplet based embeddings. *Computerized Medical Imaging and Graphics*, 86, 2020. ISSN 18790771. doi: 10.1016/j.compmedimag.2020.101794.

- G. Pascual, **P. Laiz**, A. García, H. Wenzek, J. Vitrià, and S. Seguí. Time-based self-supervised learning for wireless capsule endoscopy. *Computers in Biology and Medicine*, 146:105631, 7 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.105631.

- P. Gilabert, J. Vitrià, **P. Laiz**, C. Malagelada, A. Watson, H. Wenzek, and S. Seguí. Artificial intelligence to improve polyp detection and screening time in colon capsule endoscopy. *Front Med (Lausanne)*, 1 2022. doi: 10.3389/fmed.2022.1000726.

- **P. Laiz**, J. Vitrià, P. Gilabert, H. Wenzek, C. Malagelada, A. Watson, and S. Seguí. Anatomical landmarks localization for capsule endoscopy studies. *Computerized Medical Imaging and Graphics*, 108, 2023. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2023.102243.

- A. Quindós, **P. Laiz**, J. Vitrià and S. Seguí. Self-supervised out-of-distribution detection in Wireless Capsule Endoscopy images. *Artificial Intelligence In Medicine*, 143, 2023. ISSN 0933-3657. doi: 10.1016/j.compbiomed.2023.102606

International Conferences

- **P. Laiz**, J. Vitrià, and S. Seguí. Using the triplet loss for domain adaptation in WCE. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 399–405, 10 2019. ISBN 978-1-7281-5023-9. doi: 10.1109/ICCVW.2019.00051.

Book Chapters

- I. Amaya-Rodriguez, J. Civit-Masot, F. Luna-Perejon, L. Duran-Lopez, A. Rakhlin, S. Nikolenko, S. Kondo, **P. Laiz**, J. Vitrià, S. Seguí and P. Brandao (2021). ResNet. In: J. Bernal, A. Histace, (eds) Computer-Aided Analysis of Gastrointestinal Videos. *Springer, Cham.* doi: 10.1007/978-3-030-64340-9_12

- P. Gilabert, **P. Laiz** and S. Seguí (2023). Artificial Intelligence for Vascular Lesions. In: M. Mascarenhas, H. Cardoso and G. Macedo, (eds) Artificial Intelligence in Capsule Endoscopy. *Springer.* ISBN: 9780323996471. doi: 10.1016/B978-0-323-99647-1.00012-5

Bibliography

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

D. Abati, Angelo, P. S. Calderara, and R. Cucchiara. Latent space autoregression for novelty detection. *arXiv:1807.01653*, 2018.

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(85)80012-4.

Food and Drug Administration. Ingestible telemetric gastrointestinal capsule imaging system - final class ii special controls guidance document for industry and fda, 2001. URL `https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/ingestible-telemetric-gastrointestinal-capsule-imaging-system-\final-class-ii-special-controls`.

World Health Organization. Colorectal cancer awareness month 2022, 2022. URL `https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/`.

S. Adewole, M. Yeghyayan, D. Hyatt, L. Ehsan, J. Jablonski, A. Copland, S. Syed, and D. Brown. *Deep learning methods for anatomical landmark detection in video capsule endoscopy images*. Springer, 2020.

S. Adewole, P. Fernandez, M. Yeghyayan, J. Jablonski, A. Copland, M. D. Porter, S. Syed, and D. E. Brown. Lesion2vec: Deep metric learning for few-shot multiple lesions recognition in wireless capsule endoscopy video. *CoRR, arXiv:2101.04240*, 2021.

S. N. Adler, J. Albert, P. Baltes, F. Barbaro, C. Cellier, and J. P. Charton. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders : European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy*, pages 352–376, 2015.

M. Ahmed. Video capsule endoscopy in gastroenterology. *Gastroenterology Research*, 15(2):47, 2022.

M. Aiello, C. Cavaliere, A. D'Albore, and M. Salvatore. The challenges of diagnostic imaging in the era of big data. *Journal of clinical medicine*, 8(3):316, 2019.

A. Akay and H. Hess. Deep learning: Current and emerging applications in medicine and technology. *IEEE Journal of Biomedical and Health Informatics*, 23(3):906–920, May 2019. ISSN 2168-2194. doi: 10.1109/JBHI.2019.2894713.

A. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. Albahri, A. Alamoodi, J. Bai, A. Salhi, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 2023.

L. G. Alcalá-Gonzalez, C. Malagelada, D. M. Livovsky, and F. Azpiroz. Effect of colonic distension on small bowel motility measured by jejunal high-resolution manometry. *Neurogastroenterology & Motility*, 34(9):e14351, 2022.

L. A. Alexandre, J. Casteleiro, and N. Nobre. *Polyp detection in endoscopic video using SVMs*, volume 4702 LNAI. Springer Verlag, 2007. ISBN 9783540749752. doi: 10.1007/978-3-540-74976-9_34.

S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N. T. Nguyen, D. Q. Tran, L. D. Huynh, N. Boutry, S. Rezvy, H. Chen, Y. H. Choi, A. Subramanian, V. Balasubramanian, X. W. Gao, H. Hu, Y. Liao, D. Stoyanov, C. Daul, S. Realdon, R. Cannizzaro, D. Lamarque, T. Tran-Nguyen, A. Bailey, B. Braden, J. E. East, and J. Rittscher. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical Image Analysis*, 70:102002, 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102002.

B. Allen Jr, S. E. Seltzer, C. P. Langlotz, K. P. Dreyer, R. M. Summers, N. Petrick, D. Marinac-Dabic, M. Cruz, T. K. Alkasab, R. J. Hanisch, et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 national institutes of health/rsna/acr/the academy workshop. *Journal of the American College of Radiology*, 16(9):1179–1189, 2019.

F. Altaf, S. M. S. Islam, N. Akhtar, and N. K. Janjua. Going deep in medical image analysis: Concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019. doi: 10.1109/ACCESS.2019.2929365.

A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz. An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked*, 26:100723, 2021. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2021.100723.

T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, and T. Tada. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal Endoscopy*, 89(2):357–363.e2, feb 2019. ISSN 10976779. doi: 10.1016/j.gie.2018.10.027.

T. Aoki, A. Yamada, Y. Kato, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, and S. Ishihara. Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. *Journal of gastroenterology and hepatology*, 35(7):1196–1200, 2020.

K. Astromskė, E. Peičius, and P. Astromskis. Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. *AI & SOCIETY*, 36:509–520, 2021.

O. Attallah and M. Sharkas. GASTRO-CADx: a three stages framework for diagnosing gastrointestinal diseases. *PeerJ Computer Science*, 7:1–36, mar 2021. ISSN 23765992. doi: 10.7717/peerj-cs.423.

S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi. *Big Self-Supervised Models Advance Medical Image Classification*. Institute of Electrical and Electronics Engineers Inc., jan 2021.

S. Bae and K. Yoon. Polyp detection via imbalanced learning and discriminative feature learning. *IEEE Transactions on Medical Imaging*, 34(11):2379–2393, Nov 2015. ISSN 0278-0062. doi: 10.1109/TMI.2015.2434398.

L. Bai, L. Wang, T. Chen, Y. Zhao, and H. Ren. Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset. *Electronics*, 11(17):2747, 2022.

J. A. Barkin and J. S. Barkin. Video Capsule Endoscopy: Technology, Reading, and Troubleshooting. *Gastrointestinal Endoscopy Clinics of North America*, 27(1):15–27, 2017. ISSN 1052-5157. doi: https://doi.org/10.1016/j.giec.2016.08.002.

G. Baselli, M. Codari, and F. Sardanelli. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? *European Radiology Experimental*, 4, 2020. ISSN 25099280. doi: 10.1186/s41747-020-00159-0.

C. Becker, C. M. Christoudias, and P. Fua. Domain adaptation for microscopy imaging. *IEEE Transactions on Medical Imaging*, 34(5):1125–1139, 2015. doi: 10.1109/TMI.2014.2376872.

S. Beg, T. Card, R. Sidhu, E. Wronska, K. Ragunath, H.-L. Ching, A. Koulaouzidis, D. Yung, S. Panter, M. Mcalindon, et al. The impact of reader fatigue on the accuracy of capsule endoscopy interpretation. *Digestive and Liver Disease*, 53(8):1028–1033, 2021.

A. Belle, M. A. Kon, and K. Najarian. Biomedical informatics for computer-aided decision support systems: a survey. *The Scientific World Journal*, 2013, 2013.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79, 2010. ISSN 15730565. doi: 10.1007/s10994-009-5152-4.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–27, jan 2009. ISSN 19358237. doi: 10.1561/2200000006.

J. Berens, M. Mackiewicz, and D. Bell. *Stomach, intestine, and colon tissue discriminators for wireless capsule endoscopy images*, volume 5747. SPIE, 2005. doi: 10.1117/12.594799.

A. Biniaz, R. A. Zoroofi, and M. R. Sohrabi. Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis. *Biomedical Signal Processing and Control*, 59:101897, 2020. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2020.101897.

A. P. Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8, 2017. ISSN 18694101. doi: 10.1007/s13244-016-0534-1.

M. F. Byrne and F. Donnellan. Artificial intelligence and capsule endoscopy: Is the truly "smart" capsule nearly here? *Gastrointestinal endoscopy*, 89(1):195–197, 2019.

I. CapsoVision. Capsocam system - capsovision, 2022. URL `https://capsovision.com/capsocam-system/`.

A. Caroppo, A. Leone, and P. Siciliano. Deep transfer learning approaches for bleeding detection in endoscopy images. *Computerized Medical Imaging and Graphics*, 88(April 2020):101852, mar 2021. ISSN 18790771. doi: 10.1016/j.compmedimag.2020.101852.

H. P. Chan, L. M. Hadjiiski, and R. K. Samala. Computer-aided diagnosis in the era of deep learning, 2020a. ISSN 00942405.

H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou. Deep Learning in Medical Image Analysis. *Advances in Experimental Medicine and Biology*, 1213:3–21, 2020b. ISSN 22148019. doi: 10.1007/978-3-030-33128-3_1.

G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010.

H. Chen, J. Chen, Q. Peng, G. Sun, and T. Gan. *Automatic hookworm image detection for wireless capsule endoscopy using hybrid color gradient and contourlet transform*. Institute of Electrical and Electronics Engineers Inc., Dec 2013. doi: 10.1109/BMEI.2013.6746918.

H. Chen, X. Wu, G. Tao, and Q. Peng. Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos. *Neurocomputing*, 229:77–87, 2017a. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2016.06.077. Advances in computing techniques for big medical image data.

H. Chen, J. Cao, and M. Yi. *Out of distribution detection for medical images*, volume 12613. Springer, 2023.

Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. *Deep domain adaptation for describing people based on fine-grained clothing attributes*. Institute of Electrical and Electronics Engineers Inc., 2015.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3(Figure 1):1575–1585, 2020a.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *ACM Digital Library, NeurIPS*, pages 1–18, 2020b. ISSN 23318422.

W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:1320–1329, 2017b. doi: 10.1109/CVPR.2017.145.

W. Chen, Y. Liu, W. Wang, E. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew. Deep Image Retrieval: A Survey. *arXiv*, pages 1–20, 2021.

Y. Chen, C. Yang, and Y. Zhang. Deep domain similarity Adaptation Networks for across domain classification. *Pattern Recognition Letters*, 112:270–276, 2018. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2018.08.006.

D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. *Person re-identification by multi-channel parts-based cnn with improved triplet loss function*. Institute of Electrical and Electronics Engineers Inc., 2016.

V. Cheplygina, M. de Bruijne, and J. P. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, may 2019. ISSN 13618423. doi: 10.1016/j.media.2019.03.009.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. Association for Computational Linguistics, Doha, Qatar, Oct. 2014. doi: 10.3115/v1/D14-1179.

S. Chopra, R. Hadsell, and Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification*, volume 1. Institute of Electrical and Electronics Engineers Inc., 2005. doi: 10.1109/CVPR.2005.202.

X. Chu, C. K. Poh, L. Li, K. L. Chan, S. Yan, W. Shen, T. M. Htwe, J. Liu, J. H. Lim, E. H. Ong, et al. *Epitomized summarization of wireless capsule endoscopic videos for efficient visualization*. Springer, 2010.

E. Ciaccio, C. A Tennyson, G. Bhagat, S. Lewis, and P. Green. Implementation of a polling protocol for predicting celiac disease in videocapsule analysis. *World journal of gastrointestinal endoscopy*, 5:313–22, 07 2013. doi: 10.4253/wjge.v5.i7.313.

E. J. Ciaccio, C. A. Tennyson, S. K. Lewis, S. Krishnareddy, G. Bhagat, and P. H. Green. Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images. *Computer methods and programs in biomedicine*, 100(1):39–48, 2010.

G. Ciuti, A. Menciassi, and P. Dario. Capsule endoscopy: From current achievements to open challenges. *IEEE Reviews in Biomedical Engineering*, 4:59–72, 2011. doi: 10.1109/RBME.2011.2171182.

P. Cortegoso Valdivia, U. Deding, T. Bjørsum-Meyer, G. Baatrup, I. Fernández-Urién, X. Dray, P. Boal-Carvalho, P. Ellul, E. Toth, E. Rondonotti, et al. Inter/intra-observer agreement in video-capsule endoscopy: Are we getting it all wrong? a systematic review and meta-analysis. *Diagnostics*, 12(10):2400, 2022.

Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:9260–9269, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00949.

J. Darrow. Capsule endoscopy instead of colonoscopy? the fda approves the pillcam colon, 2014. `https://blog.petrieflom.law.harvard.edu/2014/03/04/capsule-endoscopy-instead-of-colonoscopy-the-\fda-approves-the-pillcam-colon/`.

W. Deng, L. Zheng, and J. Jiao. Domain alignment with triplets. *arXiv preprint arXiv:1812.00893*, 2018.

D. Devi, S. K. Biswas, and B. Purkayastha. *A Review on Solution to Class Imbalance Problem: Undersampling Approaches*. Institute of Electrical and Electronics Engineers Inc., 2020. doi: 10.1109/ComPE49325.2020.9200087.

G. Dharani, N. Nair, P. Satpathy, and J. Christopher. *Covariate Shift: A Review and Analysis on Classifiers*. Institute of Electrical and Electronics Engineers Inc., 2019. doi: 10.1109/GCAT47503.2019.8978471.

S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2015.04.005.

Z. Ding, H. Shi, H. Zhang, L. Meng, M. Fan, C. Han, K. Zhang, F. Ming, X. Xie, H. Liu, et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology*, 157(4):1044–1054, 2019.

H. Dokoutsidou, S. Karagiannis, E. Giannakoulopoulou, P. Galanis, N. Kyriakos, C. Liatsos, S. Faiss, and C. Mavrogiannis. A study comparing an endoscopy nurse and an endoscopy physician in capsule endoscopy interpretation. *European journal of gastroenterology & hepatology*, 23(2):166–170, 2011.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*, volume 32 of *Proceedings of Machine Learning Research*. PMLR, Bejing, China, 22–24 Jun 2014.

Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019. doi: 10.1109/ACCESS.2019.2929258.

X. Dray, C. Li, J. Saurin, F. Cholet, G. Rahmi, J. Le Mouel, C. Leandri, S. Lecleire, X. Amiot, J. Delvaux, C. Duburque, R. Gérard, R. Leenhardt, F. Mesli, G. Vanbiervliet, I. Nion Larmurier, S. Sacher-Huvelin, C. Simon-Chane, R. Olivier, and A. Histace. *CAD-CAP: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle*, volume 50. Georg Thieme Verlag KG, feb 2018. doi: 10.1055/s-0038-1623358.

X. Dray, D. Iakovidis, C. Houdeville, R. Jover, D. Diamantis, A. Histace, and A. Koulaouzidis. Artificial intelligence in small bowel capsule endoscopy - current status, challenges and future promise. *Journal of Gastroenterology and Hepatology (Australia)*, 36, 2021. ISSN 14401746. doi: 10.1111/jgh.15341.

M. Drozdzal, S. Seguí, J. Vitrià, C. Malagelada, F. Azpiroz, and P. Radeva. Adaptable image cuts for motility inspection using wce. *Computerized Medical Imaging and Graphics*, 37 (1):72–80, 2013.

M. El Ansari and S. Charfi. *Computer-aided system for polyp detection in wireless capsule endoscopy images*. Institute of Electrical and Electronics Engineers Inc., 2017.

A. El Khatib, N. Werghi, and H. Al-Ahmad. *Automatic polyp detection: A comparative study*. Institute of Electrical and Electronics Engineers Inc., Aug 2015. doi: 10.1109/EMBC.2015.7318941.

W. El-Matary. Wireless capsule endoscopy: Indications, limitations, and future challenges. *Journal of Pediatric Gastroenterology and Nutrition*, 46, 2008. ISSN 02772116. doi: 10.1097/01.mpg.0000304447.69305.cc.

R. J. Ellis, R. M. Sander, and A. Limon. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 6:100068, 2022. ISSN 2666-5212. doi: https://doi.org/10.1016/j.ibmed.2022.100068.

R. A. Enns, L. Hookey, D. Armstrong, C. N. Bernstein, C. T. Steven J Heitman 5, G. I. Leontiadis, F. Tse, and D. Sadowski. Clinical Practice Guidelines for the Use of Video

Capsule Endoscopy. *Gastroenterology*, 152(3):497–514, 2017. ISSN 0016-5085. doi: https://doi.org/10.1053/j.gastro.2016.12.032.

L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39 (3):42–62, 2022. doi: 10.1109/MSP.2021.3134634.

W. Falcon and K. Cho. A Framework For Contrastive Self-Supervised Learning And Designing A New Approach. *arXiv*, aug 2020.

Z. Falin, L. Haihua, and P. Ning. Gastrointestinal polyps and tumors detection based on multi-scale feature-fusion with wce sequences. *arXiv preprint arXiv:2204.01012*, 2022.

A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation, 2021.

Z. Feng, C. Xu, and D. Tao. Self-supervised representation learning by rotation feature decoupling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01061.

K. R. M. Fernando and C. P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2022. doi: 10.1109/TNNLS.2020.3047335.

M. Freitas, C. Arieira, P. B. Carvalho, B. Rosa, M. J. Moreira, and J. Cotter. Simplify to improve in capsule endoscopy – TOP 100 is a swift and reliable evaluation tool for the small bowel inflammatory activity in Crohn's disease. *Scandinavian Journal of Gastroenterology*, 55(4):408–413, 2020. doi: 10.1080/00365521.2020.1745880.

Y. Fu, W. Zhang, M. Mandal, and M. Q. Meng. Computer-aided bleeding detection in WCE video. *IEEE Journal of Biomedical and Health Informatics*, 18:636–642, 2014. ISSN 21682194. doi: 10.1109/JBHI.2013.2257819.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980. ISSN 1432-0770. doi: 10.1007/BF00344251.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

A. Garbaz, S. Lafraxo, S. Charfi, M. El Ansari, and L. Koutti. *Bleeding classification in wireless capsule endoscopy images based on inception-resnet-v2 and cnns*. IEEE, 2022.

F. Garcea, A. Serra, F. Lamberti, and L. Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.106391.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019. URL `https://arxiv.org/abs/1901.09192`.

L. Geng, S. Zhang, J. Tong, and Z. Xiao. Lung segmentation method with dilated convolution based on VGG-16 network. *Computer Assisted Surgery*, 24(sup2):27–33, 2019. doi: 10.1080/24699322.2019.1649071.

O. V. Geoffrey Hinton and J. Dean. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*, 3 2015. doi: 10.48550/arxiv.1503.02531.

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv*, 3 2018. doi: 10.48550/arxiv.1803.07728.

P. Gilabert, J. Vitrià, P. Laiz, C. Malagelada, A. Watson, H. Wenzek, and S. Segui. Artificial intelligence to improve polyp detection and screening time in colon capsule endoscopy. *Frontiers in Medicine*, 9, 2022. ISSN 2296-858X. doi: 10.3389/fmed.2022.1000726.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014a.

I. Goodfellow, B. Yoshua, and A. Courville. *Deep Learning*. MIT Press, 2016.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 12 2014b. doi: 10.48550/arxiv.1412.6572.

L. Goran, A. M. Negreanu, A. Stemate, and L. Negreanu. Capsule endoscopy: Current status and role in crohn's disease. *World journal of gastrointestinal endoscopy*, 10(9):184, 2018.

A. Gordo, J. Almazán, J. Revaud, and D. Larlus. *Deep image retrieval: Learning global representations for image search*. Springer, 2016.

A. Goyal, J. Kaur, J. Dhatarwal, P. Handa, and N. Goel. *Automatic detection of WCE bleeding frames using hybrid features and machine learning algorithms*. IEEE, 2022.

R. Grosse. Lecture 15: Exploding and vanishing gradients. *University of Toronto Computer Science*, 2017.

Y. Gu, Z. Ge, C. P. Bonnington, and J. Zhou. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1379–1393, 2020. doi: 10.1109/JBHI.2019.2942429.

H. Guan and M. Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022. doi: 10.1109/TBME.2021.3117407.

X. Guo and Y. Yuan. *Triple ANet: Adaptive Abnormal-aware Attention Network for WCE Image Classification*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-32239-7.

X. Guo and Y. Yuan. Semi-supervised WCE image classification with adaptive aggregated attention. *Medical Image Analysis*, 64:101733, aug 2020. ISSN 13618423. doi: 10.1016/j.media.2020.101733.

X. Guo, Z. Chen, J. Liu, and Y. Yuan. Non-equivalent images and pixels: confidence-aware resampling with meta-learning mixup for polyp segmentation. *Medical Image Analysis*, 78:102394, may 2022. ISSN 13618415. doi: 10.1016/j.media.2022.102394.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1735–1742, 2006. ISSN 10636919. doi: 10.1109/CVPR.2006.100.

M. Häfner. Conventional colonoscopy: Technique, indications, limits. *European Journal of Radiology*, 61(3):409–414, 2007. ISSN 0720048X. doi: 10.1016/j.ejrad.2006.07.034.

M. Hajabdollahi, R. Esfandiarpoor, S. M. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian. Segmentation of bleeding regions in wireless capsule endoscopy images an approach for inside capsule video summarization. *CoRR*, abs/1802.07788, 2018.

O. Haji-Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, and N. Haji-Maghsoodi. Automatic organs' detection in WCE. *AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing*, pages 116–121, 2012. doi: 10.1109/AISP.2012.6313729.

M. F. Haque, H.-Y. Lim, and D.-S. Kang. *Object Detection Based on VGG with ResNet Network*. Institute of Electrical and Electronics Engineers Inc., 2019. doi: 10.23919/ELINFOCOM.2019.8706476.

J. Hausmann, J.-P. Linke, J. G. Albert, J. Masseli, A. Tal, A. Kubesch, N. Filmann, M. Philipper, and M. Farnbacher. Time-saving polyp detection in colon capsule endoscopy: evaluation of a novel software algorithm. *International Journal of Colorectal Disease*, 34(11):1857–1863, 2019. ISSN 1432-1262. doi: 10.1007/s00384-019-03393-0.

K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*, volume 2016-Decem. Institute of Electrical and Electronics Engineers Computer Society, dec 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *Institute of Electrical and Electronics Engineers Inc.*, 11 2020. doi: 10.48550/arxiv.1911.05722.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv:1610.02136*, 2016.

D. Hendrycks, M. Mazeika, and T. G. Dietterich. Deep anomaly detection with outlier exposure. *arXiv:1812.04606*, 2018.

A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

T. Higaki, Y. Nakamura, F. Tatsugami, T. Nakaura, and K. Awai. Improvement of image quality at CT and MRI using deep learning. *Japanese Journal of Radiology*, 37(1):73–80, 2019. ISSN 1867-108X. doi: 10.1007/s11604-018-0796-2.

G. E. Hinton, A. Krizhevsky, and S. D. Wang. *Transforming auto-encoders*, volume 6791 LNCS. Springer, Berlin, Heidelberg, 2011. ISBN 9783642217340. doi: 10.1007/978-3-642-21735-7_6.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9: 1735–1780, 11 1997. ISSN 08997667. doi: 10.1162/NECO.1997.9.8.1735.

E. Hoffer and N. Ailon. *Deep metric learning using triplet network*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24261-3.

S. A. Hoogenboom, U. Bagci, and M. B. Wallace. AI in gastroenterology. The current state of play and the potential. How will it affect our practice and when? *Techniques in Gastrointestinal Endoscopy*, 22(2):42–47, 2019. ISSN 15585050. doi: 10.1016/j.tgie.2019.150634.

Y. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. *arXiv:2002.11297*, 2020.

G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks, 8 2017a. URL `https://arxiv.org/abs/1608.06993v5`.

J. Huang, R. S. Feris, Q. Chen, and S. Yan. *Cross-Domain Image Retrieval With a Dual Attribute-Aware Ranking Network*. Institute of Electrical and Electronics Engineers Inc., December 2015.

Y. Huang, H. Zheng, C. Liu, X. Ding, and G. K. Rohde. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1625–1632, 2017b. doi: 10.1109/JBHI.2017.2691738.

D. T. Huff, A. J. Weisman, and R. Jeraj. Interpretation and visualization techniques for deep learning models in medical imaging, 2021. ISSN 13616560.

Y. Hwang, J. Park, Y. J. Lim, and H. J. Chun. Application of artificial intelligence in capsule endoscopy: where are we now? *Clinical endoscopy*, 51(6):547, 2018.

D. K. Iakovidis and A. Koulaouzidis. Software for enhanced video capsule endoscopy: challenges for essential progress. *Nature Reviews Gastroenterology & Hepatology*, 12(3): 172–186, 2015.

D. K. Iakovidis, E. Spyrou, and D. Diamantis. *Efficient homography-based video visualization for wireless capsule endoscopy.* Institute of Electrical and Electronics Engineers Inc., 2013.

D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE Transactions on Medical Imaging*, 37(10):2196–2210, oct 2018. ISSN 1558254X. doi: 10.1109/TMI.2018.2837002.

G. Iddan, G. Meron, A. Glukhovsky, and P. Swain. Wireless Capsule Endoscopy. *Nature*, 405(6785):417–418, may 2000. ISSN 00280836. doi: 10.1038/35013140.

S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics*, 35, 2016. doi: 10.1145/2897824.2925974.

H. Inoue, H. Kashida, S. Kudo, M. Sasako, H. W. T. Shimoda, S. Yoshida, M. Guelrud, C. Lightdale, K. K. Wang, D. R. Riddell, R. Lambert, J. Rey, M. Jung, H. Neuhaus, A. Axon, and J. G. R. Genta. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002. *Gastrointestinal endoscopy*, 58(6 Suppl):S3–43, dec 2003. ISSN 0016-5107 (Print). doi: 10.1016/s0016-5107(03)02159-x.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*, 2 2015.

S. Jain, A. Seal, A. Ojha, O. Krejcar, J. Bureš, I. Tachecí, and A. Yazidi. Detection of abnormality in wireless capsule endoscopy images using fractal features. *Computers in Biology and Medicine*, 127:104094, dec 2020. ISSN 18790534. doi: 10.1016/j.compbiomed.2020.104094.

S. Jain, A. Seal, A. Ojha, A. Yazidi, J. Bures, I. Tacheci, and O. Krejcar. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Computers in Biology and Medicine*, 137:104789, oct 2021. ISSN 18790534. doi: 10.1016/j.compbiomed.2021.104789.

X. Jia and M. Q. . Meng. *A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images*. Institute of Electrical and Electronics Engineers Inc., Aug 2016. doi: 10.1109/EMBC.2016.7590783.

D. Jin, E. Sergeeva, W. Weng, G. Chauhan, and P. Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14, 2022. ISSN 2692-9368. doi: 10.1002/wsbm.1548.

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang. *A Comparative Study on Transformer vs RNN in Speech Applications*. Institute of Electrical and Electronics Engineers Inc., 2019. doi: 10.1109/ASRU46091.2019.9003750.

A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling, 2018.

B. Khaertdinov, E. Ghaleb, and S. Asteriadis. *Deep Triplet Networks with Attention for Sensor-based Human Activity Recognition*. Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/PERCOM50583.2021.9439116.

M. A. Khan, S. Kadry, M. Alhaisoni, Y. Nam, Y. Zhang, V. Rajinikanth, and M. S. Sarfraz. Computer-Aided Gastrointestinal Diseases Analysis from Wireless Capsule Endoscopy: A Framework of Best Features Selection. *IEEE Access*, 8:132850–132859, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3010448.

H. J. Kim, E. J. Gong, C. S. Bang, J. J. Lee, K. T. Suk, and G. H. Baik. Computer-aided diagnosis of gastrointestinal protruded lesions using wireless capsule endoscopy: A systematic review and diagnostic test accuracy meta-analysis. *Journal of Personalized Medicine*, 12(4):644, 2022.

S. H. Kim and Y. J. Lim. *Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges*, volume 11. Multidisciplinary Digital Publishing Institute, sep 2021. doi: 10.3390/diagnostics11091722.

D. P. Kingma and M. Welling. *Auto-encoding variational bayes*. Institute of Electrical and Electronics Engineers Inc., 2014.

E. Klang, Y. Barash, R. Y. Margalit, S. Soffer, O. Shimon, A. Albshesh, S. Ben-Horin, M. M. Amitai, R. Eliakim, and U. Kopylov. Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy. *Gastrointestinal endoscopy*, 91(3): 606–613, 2020.

M. Kobaek-Larsen, R. Kroijer, A. K. Dyrvig, M. M. Buijs, R. J. Steele, N. Qvist, and G. Baatrup. Back-to-back colon capsule endoscopy and optical colonoscopy in colorectal

cancer screening individuals. *Colorectal Disease*, 20(6):479–485, 2018. ISSN 14631318. doi: 10.1111/codi.13965.

A. Koffas, F.-m. Laskaratos, and O. Epstein. Non-small bowel lesion detection at small bowel capsule endoscopy: A comprehensive literature review. *World Journal of Clinical Cases*, 6(15):901–907, 2018. doi: 10.12998/wjcc.v6.i15.901.

L. Kohoutová, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, and C.-W. Woo. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, 15(4):1399–1435, 2020. ISSN 1750-2799. doi: 10.1038/s41596-019-0289-5.

A. Koulaouzidis, K. Dabos, M. Philipper, E. Toth, and M. Keuchel. How should we do colon capsule endoscopy reading: a practical guide. *Therapeutic Advances in Gastrointestinal Endoscopy*, 9(6):259–261, 2021. ISSN 2040-6207. doi: 10.1177/https.

A. Koulaouzidis, T. Bjørsum, and E. Toth. Real-life practice data on colon capsule endoscopy: We need them fast! *Endoscopy international open*, 10(03):E230–E231, 2022.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 5 2012. ISSN 15577317. doi: 10.1145/3065386.

B. Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013. ISSN 1935-8237. doi: 10.1561/2200000019.

R. Kumar, Q. Zhao, S. Seshamani, G. Mullin, G. Hager, and T. Dassopoulos. Assessment of crohn's disease lesions in wireless capsule endoscopy images. *Biomedical Engineering, IEEE Transactions on*, 59(2):355–362, Feb 2012. ISSN 0018-9294.

A. K. Kundu and S. A. Fattah. Probability density function based modeling of spatial feature variation in capsule endoscopy data for automatic bleeding detection. *Computers in Biology and Medicine*, 115:103478, dec 2019. ISSN 18790534. doi: 10.1016/j.compbiomed.2019.103478.

N. Kurniawan and M. Keuchel. *Video Capsule Endoscopy, Technology.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-662-44062-9. doi: 10.1007/978-3-662-44062-9_3.

P. Laiz, J. Vitria, and S. Segui. *Using the Triplet Loss for Domain Adaptation in WCE.* Institute of Electrical and Electronics Engineers Inc., oct 2019. ISBN 978-1-7281-5023-9. doi: 10.1109/ICCVW.2019.00051.

P. Laiz, J. Vitrià, H. Wenzek, C. Malagelada, F. Azpiroz, and S. Seguí. WCE polyp detection with triplet based embeddings. *Computerized Medical Imaging and Graphics*, 86, 2020. ISSN 18790771. doi: 10.1016/j.compmedimag.2020.101794.

P. Laiz, J. Vitrià, P. Gilabert, H. Wenzek, C. Malagelada, A. J. Watson, and S. Seguí. Anatomical landmarks localization for capsule endoscopy studies. *Computerized Medical Imaging and Graphics*, 108:102243, 2023. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2023.102243.

L. Lan and C. Ye. Recurrent generative adversarial networks for unsupervised WCE video summarization. *Knowledge-Based Systems*, 222, 2021. ISSN 00942405. doi: 10.1118/1.4816310.

L. Lan, C. Ye, C. Wang, and S. Zhou. Deep convolutional neural networks for wce abnormality detection: Cnn architecture, region proposal and transfer learning. *IEEE Access*, 7:30017–30032, 2019.

J. Latif, C. Xiao, A. Imran, and S. Tu. *Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review*. Institute of Electrical and Electronics Engineers Inc., 2019. doi: 10.1109/ICOMET.2019.8673502.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

H. Y. Lee, J. B. Huang, M. Singh, and M. H. Yang. *Unsupervised Representation Learning by Sorting Sequences*, volume 2017-Octob. Institute of Electrical and Electronics Engineers Inc., 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.79.

J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang. *Automatic Classification of Digestive Organs in Wireless Capsule Endoscopy Videos*. SAC '07. Association for Computing Machinery, New York, NY, USA, 2007. ISBN 1595934804. doi: 10.1145/1244002.1244230.

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *arXiv:1807.03888*, 2018.

R. Leenhardt, P. Vasseur, C. Li, J. C. Saurin, G. Rahmi, F. Cholet, A. Becq, P. Marteau, A. Histace, X. Dray, et al. A neural network algorithm for detection of gi angiectasia during small-bowel capsule endoscopy. *Gastrointestinal endoscopy*, 89(1):189–194, 2019.

I. I. Lei, G. J. Nia, E. White, H. Wenzek, S. Segui, A. J. Watson, A. Koulaouzidis, and R. P. Arasaradnam. Clinicians' guide to artificial intelligence in colon capsule endoscopy—technology made simple. *Diagnostics*, 13(6):1038, 2023a.

I. I. Lei, K. Tompkins, E. White, A. Watson, N. Parsons, A. Noufaily, S. Segui, H. Wenzek, R. Badreldin, A. Conlin, and R. P. Arasaradnam. Study of capsule endoscopy delivery at scale through enhanced artificial intelligence-enabled analysis (the CESCAIL study). *Colorectal Disease*, 2023b. doi: https://doi.org/10.1111/codi.16575.

B. S. Lewis, G. M. Eisen, and S. Friedman. A pooled analysis to evaluate results of capsule endoscopy trials. *Endoscopy*, 37, 2005. ISSN 0013726X. doi: 10.1055/s-2005-870353.

B. Li and M. Q.-H. Meng. Automatic polyp detection for wireless capsule endoscopy images. *Expert Systems with Applications*, 39:10952–10958, 09 2012. doi: 10.1016/j.eswa.2012.03.029.

B. Li, Y. Fan, M. Q. . Meng, and L. Qi. *Intestinal polyp recognition in capsule endoscopy images using color and shape features.* Institute of Electrical and Electronics Engineers Inc., Dec 2009. doi: 10.1109/ROBIO.2009.5420969.

B. Li, G. Xu, R. Zhou, and T. Wang. Computer aided wireless capsule endoscopy video segmentation. *Medical Physics*, 42(2):645–652, 2015. ISSN 00942405. doi: 10.1118/1.4905164.

D. Li, Y. Yang, Y. Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization, 2017. ISSN 15505499.

J. Li and W. K. Leung. Colon capsule endoscopy for inflammatory bowel disease. *Journal of Digestive Diseases*, 19(7):386–394, 2018. ISSN 17512980. doi: 10.1111/1751-2980.12614.

P. Li, Z. Li, F. Gao, L. Wan, and J. Yu. *Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images.* Institute of Electrical and Electronics Engineers Inc., July 2017. doi: 10.1109/ICME.2017.8019415.

X. Li, Y. Lu, C. Desrosiers, and X. Liu. Out-of distribution detection for skin lesion images with deep isolation forest. *arXiv:2003.09365*, 2020.

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv:1706.02690*, 2017.

Y. Liao, R. Huang, J. Li, Z. Chen, and W. Li. Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed. *IEEE Transactions on Instrumentation and Measurement*, 69(10):8064–8075, 2020.

M. Liedlgruber and A. Uhl. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. *IEEE Reviews in Biomedical Engineering*, 4, 2011. ISSN 19373333. doi: 10.1109/RBME.2011.2175445.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, jun 2021. ISSN 15582191. doi: 10.1109/TKDE.2021.3090866.

X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35 (1):857–876, 2023. doi: 10.1109/TKDE.2021.3090866.

S.-C. B. Lo, J.-S. Lin, M. T. F. M.D., and S. K. Mun. *Computer-assisted diagnosis of lung nodule detection using artificial convolution neural network*, volume 1898. SPIE, 1993. doi: 10.1117/12.154572.

S.-C. B. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7): 1201–1214, 1995. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(95)00061-5.

M. Long, Z. CAO, J. Wang, and M. I. Jordan. *Conditional Adversarial Domain Adaptation*, volume 31. Curran Associates, Inc., 2018.

I. Lopes, A. Silva, M. Coimbra, M. Dinis-Ribeiro, D. Libânio, and F. Renna. *Supervised and semi-supervised training of deep convolutional neural networks for gastric landmark detection.* IEEE, 2022.

Y. Lu and P. Xu. Anomaly detection for skin disease images using variational autoencoder. *arXiv:1807.01349*, 2018.

A. S. Lundervold and A. Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019. ISSN 0939-3889. doi: https://doi.org/10.1016/j.zemedi.2018.11.002.

G. Lv, G. Yan, and Z. Wang. *Bleeding detection in wireless capsule endoscopy images based on color invariants and spatial pyramids using support vector machines*, volume 2011. Annu Int Conf IEEE Eng Med Biol Soc, 2011. ISBN 9781424441211. doi: 10.1109/IEMBS.2011.6091638.

G. M Cobrin, R. H Pittman, and B. Lewis. Increased diagnostic yield of small bowel tumors with capsule endoscopy. *Cancer*, 107:22–7, 07 2006. doi: 10.1002/cncr.21975.

M. A. M. Meseeha. Colon polyps - statpearls publishing, 2022. URL `https://www.ncbi.nlm.nih.gov/books/NBK430761/`.

L. Ma, M. M. Crawford, L. Zhu, and Y. Liu. Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2305–2323, 2019. doi: 10.1109/TGRS.2018.2872850.

M. Ma, H. Fan, and K. M. Kitani. *Going deeper into first-person activity recognition.* Institute of Electrical and Electronics Engineers Inc., 2016.

M. Mackiewicz, J. Berens, and M. Fisher. Wireless capsule endoscopy color video segmentation. *IEEE Transactions on Medical Imaging*, 27(12):1769–1781, 2008. ISSN 02780062. doi: 10.1109/TMI.2008.926061.

A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood. *Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation.* Institute of Electrical and Electronics Engineers Inc., 2018. doi: 10.1109/ISBI.2018.8363749.

Y. Maeda, S.-e. Kudo, Y. Mori, M. Misawa, N. Ogata, S. Sasanuma, K. Wakamura, M. Oda, K. Mori, and K. Ohtsuka. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). *Gastrointestinal endoscopy*, 89(2):408–415, 2019.

A. Maieron, D. Hubner, B. Blaha, C. Deutsch, T. Schickmair, A. Ziachehabi, E. Kerstan, P. Knoflach, and R. Schoefl. Multicenter retrospective evaluation of capsule endoscopy in clinical routine. *Endoscopy*, 36(10):864–868, 2004.

C. Malagelada, F. De Iorio, F. Azpiroz, A. Accarino, S. Segui, P. Radeva, and J. R. Malagelada. New Insight Into Intestinal Motor Function via Noninvasive Endoluminal Image Analysis. *Gastroenterology*, 135(4):1155–1162, 2008. ISSN 00165085. doi: 10.1053/j.gastro.2008.06.084.

C. Malagelada, S. Seguí, S. Mendez, M. Drozdzal, J. Vitria, P. Radeva, J. Santos, A. Accarino, J. Malagelada, F. Azpiroz, et al. Functional gut disorders or disordered gut function? small bowel dysmotility evidenced by an original technique. *Neurogastroenterology & Motility*, 24(3):223–e105, 2012.

C. Malagelada, M. Drozdzal, S. Seguí, S. Méndez, J. Vitrià, P. Radeva, J. Santos, A. Accarino, J. R Malagelada, and F. Azpiroz. Classification of functional bowel disorders by objective physiological criteria based on endoluminal image analysis. *American journal of physiology. Gastrointestinal and liver physiology*, 309:ajpgi.00193.2015, 08 2015. doi: 10.1152/ajpgi.00193.2015.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

J. J. McGoran, M. E. McAlindon, P. G. Iyer, E. J. Seibel, R. Haidry, L. B. Lovat, and S. S. Sami. Miniature gastrointestinal endoscopy: Now and the future. *World Journal of Gastroenterology*, 25(30):4051–4060, 2019. ISSN 22192840. doi: 10.3748/wjg.v25.i30.4051.

L. McInnes, J. Healy, N. Saul, and L. Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

C. M. McLeavy, M. H. Chunara, R. J. Gravell, A. Rauf, A. Cushnie, C. Staley Talbot, and R. M. Hawkins. The future of CT: deep learning reconstruction. *Clinical Radiology*, 76 (6):407–415, 2021. ISSN 0009-9260. doi: https://doi.org/10.1016/j.crad.2021.01.010.

J. Mi, X. Han, R. Wang, R. Ma, and D. Zhao. Diagnostic accuracy of wireless capsule endoscopy in polyp recognition using deep learning: A meta-analysis. *International Journal of Clinical Practice*, 2022, 2022.

A. Mikołajczyk and M. Grochowski. *Data augmentation for improving deep learning in image classification problem*. Institute of Electrical and Electronics Engineers Inc., May 2018. doi: 10.1109/IIPHDW.2018.8388338.

I. Misra and L. van der Maaten. *Self-supervised learning of pretext-invariant representations*. IEEE Computer Society, dec 2020. doi: 10.1109/CVPR42600.2020.00674.

I. Misra, C. Lawrence Zitnick, and M. Hebert. *Shuffle and learn: Unsupervised learning using temporal order verification*, volume 9905 LNCS. Springer, Cham, 2016. ISBN 9783319464473. doi: 10.1007/978-3-319-46448-0_32.

M. Mostapha and M. Styner. Role of deep learning in infant brain MRI analysis. *Magnetic Resonance Imaging*, 64:171–189, 2019. ISSN 0730-725X. doi: https://doi.org/10.1016/j.mri.2019.06.009.

P. Muruganantham and S. M. Balakrishnan. A survey on deep learning models for wireless capsule endoscopy image analysis. *International Journal of Cognitive Computing in Engineering*, 2:83–92, 2021. ISSN 2666-3074. doi: https://doi.org/10.1016/j.ijcce.2021.04.002.

P. Muruganantham and S. M. Balakrishnan. Attention Aware Deep Learning Model for Wireless Capsule Endoscopy Lesion Classification and Localization. *Journal of Medical and Biological Engineering*, 42(2):157–168, 2022. ISSN 2199-4757. doi: 10.1007/s40846-022-00686-8.

E. S. Nadimi, M. M. Buijs, J. Herp, R. Kroijer, M. Kobaek-Larsen, E. Nielsen, C. D. Pedersen, V. Blanes-Vidal, and G. Baatrup. Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy. *Computers and Electrical Engineering*, 81:106531, jan 2020. ISSN 00457906. doi: 10.1016/j.compeleceng.2019.106531.

N. G. Nair, P. Satpathy, J. Christopher, et al. *Covariate shift: A review and analysis on classifiers*. IEEE, 2019.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines, 2010.

F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, and B. H. Menze. Evaluating the Robustness of Self-Supervised Learning in Medical Imaging. *arXiv*, may 2021.

H. Noh, T. Kim, J. Mun, and B. Han. *Transfer learning via unsupervised task discovery for visual question answering*. Institute of Electrical and Electronics Engineers Inc., 2019.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. *Learning and transferring mid-level image representations using convolutional neural networks*. Institute of Electrical and Electronics Engineers Inc., 2014.

K. Ostherr. Artificial intelligence and medical humanities. *Journal of Medical Humanities*, 43(2):211–232, 2022.

T. Ozawa, S. Ishihara, M. Fujishiro, H. Saito, Y. Kumagai, S. Shichijo, K. Aoyama, and T. Tada. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointestinal endoscopy*, 89(2):416–421, 2019.

N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. *Image Transformer*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 10–15 Jul 2018.

G. Pascual, P. Laiz, A. García, H. Wenzek, J. Vitrià, and S. Seguí. Time-based self-supervised learning for Wireless Capsule Endoscopy. *Computers in Biology and Medicine*, 146:105631, 2022a. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105631.

G. Pascual, J. Vitrià, and S. Seguí. *Time-coherent embeddings for Wireless Capsule Endoscopy*. Institute of Electrical and Electronics Engineers Inc., 2022b. doi: 10.1109/ICPR56361.2022.9956652.

D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. *Learning features by watching objects move*, volume 2017-Janua. Institute of Electrical and Electronics Engineers Inc., dec 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.638.

F. Pérez-García, R. Dorent, M. Rizzi, F. Cardinale, V. Frazzini, V. Navarro, C. Essert, I. Ollivier, T. Vercauteren, R. Sparks, J. S. Duncan, and S. Ourselin. A self-supervised learning strategy for postoperative brain cavity segmentation simulating resections. *International Journal of Computer Assisted Radiology and Surgery*, 82:1–9, jun 2021. ISSN 18616429. doi: 10.1007/s11548-021-02420-2.

K. Pogorelov, S. Suman, F. A. Hussin, A. S. Malik, O. Ostroukhova, M. Riegler, P. Halvorsen, S. H. Ho, and K. L. Goh. Bleeding detection in wireless capsule endoscopy videos — color versus texture features. *Journal of Applied Clinical Medical Physics*, 20:141–154, 8 2019. ISSN 15269914. doi: 10.1002/acm2.12662.

V. H. Pong, M. Dalal, S. Lin, A. Nair, S. Bahl, and S. Levine. Skew-fit: State-covering self-supervised reinforcement learning. *PMLR*, 3 2019. doi: 10.48550/arxiv.1903.03698.

A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180, 2020.

X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1647-3.

A. Quindós, P. Laiz, J. Vitrià, and S. Seguí. Self-supervised out-of-distribution detection in wireless capsule endoscopy images. *Artificial Intelligence in Medicine*, 143, 2023. ISSN 0933-3657. doi: https://doi.org/10.1016/j.compbiomed.2023.102606.

T. Rahim, M. A. Usman, and S. Y. Shin. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Computerized Medical Imaging and Graphics*, page 101767, 2020.

M. Rana and M. Bhushan. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*, pages 1–39, 2022.

V. Raut and R. Gunjan. Transfer learning based video summarization in wireless capsule endoscopy. *International Journal of Information Technology*, 14(4):2183–2190, 2022. ISSN 2511-2112. doi: 10.1007/s41870-022-00894-0.

J. Reuss, G. Pascual, H. Wenzek, and S. Seguí. Sequential Models for Endoluminal Image Classification. *Diagnostics*, 12(2), 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12020501.

T. Ribeiro, M. Mascarenhas, J. Afonso, H. Cardoso, P. Andrade, S. Lopes, J. Ferreira, M. Mascarenhas Saraiva, and G. Macedo. Artificial intelligence and colon capsule endoscopy: Automatic detection of ulcers and erosions using a convolutional neural network. *Journal of Gastroenterology and Hepatology*, 2022.

A. R. Robertson, S. Segui, H. Wenzek, and A. Koulaouzidis. Artificial intelligence for the detection of polyps or cancer with colon capsule endoscopy. *Therapeutic Advances in Gastrointestinal Endoscopy*, 14:26317745211020277, 2021.

E. Rondonotti, M. Pennazio, E. Toth, and A. Koulaouzidis. How to read small bowel capsule endoscopy: a practical guide for everyday use. *Endoscopy International Open*, 08(10): E1220–E1224, 2020. ISSN 2364-3722. doi: 10.1055/a-1210-4830.

A. G. Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver, N. Vo, P. Bui, S. Winter, P. MacWilliams, G. S. Corrado, U. Telang, Y. Liu, A. T. Cemgil, A. Karthikesalingam, B. Lakshminarayanan, and J. Winkens. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *arXiv:2104.03829*, 2021.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation (original). *Explorations in the Micro-Structure of Cognition Vol. 1 : Foundations*, 1986.

F. Rustam, M. A. Siddique, H. U. R. Siddiqui, S. Ullah, A. Mehmood, I. Ashraf, and G. S. Choi. Wireless Capsule Endoscopy Bleeding Images Classification Using CNN Based Model. *IEEE Access*, 9:33675–33688, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3061592.

H. Saito, T. Aoki, K. Aoyama, Y. Kato, A. Tsuboi, A. Yamada, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal endoscopy*, 92(1):144–151, 2020.

M. K. Sana, Z. M. Hussain, P. A. Shah, and M. H. Maqsood. Artificial intelligence in celiac disease. *Computers in Biology and Medicine*, 125:103996, 2020.

M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine*, 13(4):59–76, 2018. doi: 10.1109/MCI.2018.2866730.

V. Saraf, P. Chavan, and A. Jadhav. *Deep Learning Challenges in Medical Imaging*. Springer Singapore, Singapore, 2020. ISBN 978-981-15-3242-9.

M. M. Saraiva, J. P. Ferreira, H. Cardoso, J. Afonso, T. Ribeiro, P. Andrade, M. P. Parente, R. N. Jorge, and G. Macedo. Artificial intelligence and colon capsule endoscopy: automatic detection of blood in colon capsule endoscopy using a convolutional neural network. *Endoscopy International Open*, 9(08):E1264–E1268, 2021.

J.-C. Saurin, M. G. Lapalus, F. Cholet, P. N. D'Halluin, B. Filoche, M. Gaudric, S. Sacher-Huvelin, C. Savalle, M. Frederic, P. Adenis Lamarre, and E. Ben Soussan. Can we shorten the small-bowel capsule reading time with the "Quick-view" image detection system? *Digestive and Liver Disease*, 44(6):477–481, 2012. ISSN 1590-8658. doi: https://doi.org/10.1016/j.dld.2011.12.021.

E. Saxena, M. Yadav, M. Yadav, and P. Shoran. *Artificial Intelligence-based Diagnostic Analysis for Wireless Capsule Endoscopy in Obscure Bowel Disease Detection: A Potential*. ACM Digital Library, 2022.

D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition, 2010. ISSN 03029743. URL `https://link-springer-com.sire.ub.edu/chapter/10.1007/978-3-642-15825-4_10`.

F. Schroff, D. Kalenichenko, and J. Philbin. *FaceNet: A unified embedding for face recognition and clustering*, volume 07-12-June-2015. Institute of Electrical and Electronics Engineers Inc., 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298682.

M. Schultz and T. Joachims. Learning a distance metric from relative comparisons, 2004. ISSN 10495258.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

S. Segui, M. Drozdzal, E. Zaytseva, C. Malagelada, F. Azpiroz, P. Radeva, and J. Vitria. Detection of wrinkle frames in endoluminal videos using betweenness centrality measures

for images. *Biomedical and Health Informatics, IEEE Journal of*, PP(99):1–1, 2014. ISSN 2168-2194. doi: 10.1109/JBHI.2014.2304179.

S. Seguí, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, and J. Vitrià. Generic feature learning for wireless capsule endoscopy analysis. *Computers in Biology and Medicine*, 79:163–172, 2016. ISSN 18790534. doi: 10.1016/j.compbiomed.2016.10.011.

T. J. Sejnowski. *The deep learning revolution*. MIT press, 2018.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, Oct 2017. ISSN 2380-7504.

P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. *Time-Contrastive Networks: Self-Supervised Learning from Video*. Institute of Electrical and Electronics Engineers Inc., sep 2018. ISBN 9781538630815. doi: 10.1109/ICRA.2018.8462891.

S. Serte and H. Demirel. Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in Biology and Medicine*, 132:104306, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104306.

A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. *Adversarial training for free!*, volume 32. Curran Associates, Inc., 2019.

R. Shahril, A. Saito, A. Shimizu, and S. Baharun. Bleeding Classification of Enhanced Wireless Capsule Endoscopy Images using Deep Convolutional Neural Network. *Journal of Information Science and Engineering*, 36(1):91–108, 2020. ISSN 10162364. doi: 10.6688/JISE.20200136(1).0006.

H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations, 2016. ISSN 16113349.

A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. *Learning from Simulated and Unsupervised Images through Adversarial Training*. Institute of Electrical and Electronics Engineers Inc., 2017. doi: 10.1109/CVPR.2017.241.

A. Sieg. Capsule endoscopy compared with conventional colonoscopy for detection of colorectal neoplasms. *World Journal of Gastrointestinal Endoscopy*, 3(5):81, 2011. ISSN 1948-5190. doi: 10.4253/wjge.v3.i5.81.

R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 9 2015.

A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler, and P. Halvorsen. Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):1–10, 2021. ISSN 20524463. doi: 10.1038/s41597-021-00920-z.

L. N. Smith. *Cyclical Learning Rates for Training Neural Networks*. Institute of Electrical and Electronics Engineers Inc., March 2017. doi: 10.1109/WACV.2017.58.

M. E. Smith and D. G. Morton. *The Digestive System*. Elsevier, 2010. ISBN 978-0-7020-3367-4. doi: 10.1016/C2009-0-51554-1.

K. Sohn. Improved deep metric learning with multi-class N-pair loss objective. *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016. ISSN 10495258.

G. Son, T. Eo, J. An, D. J. Oh, Y. Shin, H. Rha, Y. J. Kim, Y. J. Lim, and D. Hwang. Small bowel detection for wireless capsule endoscopy using convolutional neural networks with temporal filtering. *Diagnostics (Basel, Switzerland)*, 12(8):1858, July 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12081858.

C. Spada, C. Hassan, D. Bellini, D. Burling, G. Cappello, C. Carretero, E. Dekker, R. Eliakim, M. de Haan, M. F. Kaminski, A. Koulaouzidis, A. Laghi, P. Lefere, T. Mang, S. M. Milluzzo, M. Morrin, D. McNamara, E. Neri, S. Pecere, M. Pioche, A. Plumb, E. Rondonotti, M. C. W. Spaander, S. Taylor, I. Fernandez-Urien, J. E. van Hooft, J. Stoker, and D. Regge. Imaging alternatives to colonoscopy: CT colonography and colon capsule. European Society of Gastrointestinal Endoscopy (ESGE) and European Society of Gastrointestinal and Abdominal Radiology (ESGAR) Guideline – Update 2020. *European Radiology*, 31(5):2967–2982, 2021. ISSN 1432-1084. doi: 10.1007/s00330-020-07413-4.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 2014. ISSN 15337928.

Y. Sun, Y. Ming, X. Zhu, and Y. Li. Out-of-distribution detection with deep nearest neighbors. *PMLR*, 4 2022.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 9 2015. ISSN 10636919. URL `https://arxiv.org/abs/1409.4842v1`.

S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh. Combo loss: Handling input and output imbalance in multi-organ seg-

mentation. *Computerized Medical Imaging and Graphics*, 75:24–33, 2019. ISSN 18790771. doi: 10.1016/j.compmedimag.2019.04.005.

F. W. D. Tai, M. McAlindon, and R. Sidhu. Colon capsule endoscopy–shining the light through the colon. *Current Gastroenterology Reports*, pages 1–7, 2023.

K. Takada, Y. Yabuuchi, and N. Kakushima. Evaluation of current status and near future perspectives of capsule endoscopy: Summary of Japan Digestive Disease Week 2019. *Digestive Endoscopy*, 32(4):529–531, 2020. ISSN 14431661. doi: 10.1111/den.13659.

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June: 10691–10700, 5 2019. doi: 10.48550/arxiv.1905.11946.

A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed. Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10:47643–47660, 2022. doi: 10.1109/ACCESS.2022.3169512.

S. Thulasidasan, T. Bhattacharya, J. Bilmes, G. Chennupati, and J. Mohd-Yusof. Combating label noise in deep learning using abstention. *arXiv:1905.10964*, 2019.

R. Trasolini and M. F. Byrne. *Artificial intelligence and deep learning for small bowel capsule endoscopy*, volume 33. John Wiley & Sons, Ltd, jan 2021. doi: 10.1111/den.13896.

M. Tschannen, J. Djolonga, M. Ritter, A. Mahendran, N. Houlsby, S. Gelly, and M. Lucic. *Self-Supervised Learning of Video-Induced Visual Invariances*. Institute of Electrical and Electronics Engineers Inc., 2015.

A. B. Tufail, Y.-K. Ma, M. K. A. Kaabar, F. Martínez, A. R. Junejo, I. Ullah, and R. Khan. Deep learning in cancer diagnosis and prognosis prediction: A minireview on challenges, recent trends, and future directions. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

T. Ueda, Y. Ohno, K. Yamamoto, K. Murayama, M. Ikedo, M. Yui, S. Hanamatsu, Y. Tanaka, Y. Obama, H. Ikeda, and H. Toyama. Deep Learning Reconstruction of Diffusion-weighted MRI Improves Image Quality for Prostatic Imaging. *Radiology*, 303 (Genitourinary Imaging), 2022.

M. A. Usman, G. Satrya, M. R. Usman, and S. Y. Shin. Detection of small colon bleeding in wireless capsule endoscopy videos. *Computerized Medical Imaging and Graphics*, 54:16 – 26, 2016. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2016.09.005.

V. V and K. V. Prashanth. Ulcer detection in Wireless Capsule Endoscopy images using deep CNN. *Journal of King Saud University - Computer and Information Sciences*, sep 2020. ISSN 22131248. doi: 10.1016/j.jksuci.2020.09.008.

A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. *Pixel recurrent neural networks*. PMLR, 2016.

A. van den Oord, Y. Li, and O. Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv*, jul 2018.

L. Van Der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2625, 2008. ISSN 15324435.

G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.

M. Vasilakakis, A. Koulaouzidis, D. E. Yung, J. N. Plevris, E. Toth, and D. K. Iakovidis. Follow-up on: optimizing lesion detection in small bowel capsule endoscopy and beyond: from present problems to future solutions. *Expert Review of Gastroenterology and Hepatology*, 13(2):129–141, 2019. ISSN 17474132. doi: 10.1080/17474124.2019.1553616.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December, 2017. ISSN 10495258.

A. Vats, M. Pedersen, A. Mohammed, and Ø. Hovde. *Learning More for Free - A Multi Task Learning Approach for Improved Pathology Classification in Capsule Endoscopy*, volume 12907 LNCS. Springer International Publishing, 2021. ISBN 9783030872335. doi: 10.1007/978-3-030-87234-2_1.

A. Vats, A. Mohammed, and M. Pedersen. From labels to priors in capsule endoscopy: a prior guided approach for improving generalization with few labels. *Scientific Reports*, 12(1):15708, 2022.

P. M. Vieira, N. R. Freitas, V. B. Lima, D. Costa, C. Rolanda, and C. S. Lima. Multi-pathology detection and lesion localization in WCE videos by using the instance segmentation approach. *Artificial Intelligence in Medicine*, 119(June):102141, 2021. ISSN 18732860. doi: 10.1016/j.artmed.2021.102141.

A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, 2018. ISSN 16875273. doi: 10.1155/2018/7068349.

A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. *arXiv:1809.03576*, 2018.

S. Waite, Z. Farooq, A. Grigorian, C. Sistrom, S. Kolla, A. Mancuso, S. Martinez-Conde, R. G. Alexander, A. Kantor, and S. L. Macknik. A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era

of artificial intelligence. *Academic Radiology*, 27(1):26–38, 2020. ISSN 1076-6332. doi: https://doi.org/10.1016/j.acra.2019.08.018. Special Issue: Artificial Intelligence.

J. Wan, B. Chen, and Y. Yu. Polyp detection from colorectum images by using attentive YOLOv5. *Diagnostics*, 11(12), 2021. ISSN 20754418. doi: 10.3390/diagnostics11122264.

F. Wang, L. P. Casalino, and D. Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294, 2019a.

J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep Metric Learning with Angular Loss. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October: 2612–2620, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.283.

J. Wang, H. Zhu, S. H. Wang, and Y. D. Zhang. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26, 2021. ISSN 15728153. doi: 10.1007/s11036-020-01672-7.

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.05.083.

S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang. A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks. *Physics in Medicine and Biology*, 64(23):235014, dec 2019b. doi: 10.1088/1361-6560/ab5086.

S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang. Deep Convolutional Neural Network for Ulcer Recognition in Wireless Capsule Endoscopy: Experimental Feasibility and Optimization. *Computational and Mathematical Methods in Medicine*, 2019:7546215, 2019c. ISSN 1748-670X. doi: 10.1155/2019/7546215.

X. Wang and A. Gupta. *Unsupervised learning of visual representations using videos*, volume 2015 Inter. Institute of Electrical and Electronics Engineers Inc., 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.320.

X. Wang and M. Q. Meng. Perspective of active capsule endoscope: actuation and localisation. *International Journal of Mechatronics and Automation*, 1, 2011. ISSN 20451067. doi: 10.1504/IJMA.2011.039154.

X. Wang and J. Schneider. *Flexible Transfer Learning under Support and Model Shift*, volume 27. Curran Associates, Inc., 2014.

X. Wang, A. Jabri, and A. A. Efros. *Learning correspondence from the cycle-consistency of time*, volume 2019-June. Institute of Electrical and Electronics Engineers Inc., 2019d. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00267.

X. Wang, H. Qian, E. J. Ciaccio, S. K. Lewis, G. Bhagat, P. H. Green, S. Xu, L. Huang, R. Gao, and Y. Liu. Celiac disease diagnosis from videocapsule endoscopy images

with residual learning and deep feature extraction. *Computer Methods and Programs in Biomedicine*, 187:105236, apr 2020. ISSN 18727565. doi: 10.1016/j.cmpb.2019.105236.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009. ISSN 15324435.

M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du, and X. Lan. A review of object detection based on deep learning. *Multimedia Tools and Applications*, 79(33):23729–23791, 2020. ISSN 1573-7721. doi: 10.1007/s11042-020-08976-6.

Z. Xiao, J. Lu, X. Wang, N. Li, Y. Wang, and N. Zhao. WCE-DCGAN: A data augmentation method based on wireless capsule endoscopy images for gastrointestinal disease detection. *IET Image Processing*, pages 1–11, 2022. ISSN 17519667. doi: 10.1049/ipr2.12704.

J. Xing and H. Mouchère. *Contrastive Self-Supervised Learning on Crohn's Disease Detection*. IEEE, 2022.

D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. *Self-supervised spatiotemporal learning via video clip order prediction*, volume 2019-June. Institute of Electrical and Electronics Engineers Inc., 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.01058.

H. Xuan, A. Stylianou, and R. Pless. Improved embeddings with easy positive triplet mining. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 2463–2471, 2020. doi: 10.1109/WACV45572.2020.9093432.

S. S. Yadav and S. M. Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):113, 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0276-2.

H. Yamamoto, H. Ogata, T. Matsumoto, N. Ohmiya, K. Ohtsuka, K. Watanabe, T. Yano, T. Matsui, K. Higuchi, T. Nakamura, and K. Fujimoto. Clinical Practice Guideline for Enteroscopy. *Digestive Endoscopy*, pages 519–546, 2017. doi: 10.1111/den.12883.

J. Yanase and E. Triantaphyllou. The seven key challenges for the future of computer-aided diagnosis in medicine. *International journal of medical informatics*, 129:413–422, 2019a.

J. Yanase and E. Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138:112821, 2019b.

J. Yang, L. Chang, S. Li, X. He, and T. Zhu. WCE polyp detection based on novel feature descriptor with normalized variance locality-constrained linear coding. *International Journal of Computer Assisted Radiology and Surgery*, 15(8):1291–1302, 2020.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection : Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–22, 2022.

Y. J. Yang. The Future of Capsule Endoscopy: The Role of Artificial Intelligence and Other Technical Advancements. *Clinical Endoscopy*, 53(4):387–394, 2020. ISSN 2234-2400. doi: 10.5946/ce.2020.133.

J.-Y. Yeh, T.-H. Wu, and W.-J. Tsai. Bleeding and Ulcer Detection Using Wireless Capsule Endoscopy Images. *Journal of Software Engineering and Applications*, 07(05):422–432, may 2014. ISSN 1945-3116. doi: 10.4236/jsea.2014.75039.

B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao. *Correcting the triplet selection bias for triplet loss*. Springer, 2018.

J. Yu, J. Chen, Z. Q. Xiang, and Y. Zou. *A hybrid convolutional neural networks with extreme learning machine for WCE image classification*. Institute of Electrical and Electronics Engineers Inc., Dec 2015. doi: 10.1109/ROBIO.2015.7419037.

L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE Journal of Biomedical and Health Informatics*, 21(1):65–75, Jan 2017. ISSN 2168-2194. doi: 10.1109/JBHI.2016.2637004.

X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, and Y.-D. Zhang. Transfer learning for medical images analyses: A survey. *Neurocomputing*, 489:230–254, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.08.159.

Y. Yuan and M. Q. . Meng. *A novel feature for polyp detection in wireless capsule endoscopy images*. Institute of Electrical and Electronics Engineers Inc., Sep. 2014. doi: 10.1109/IROS.2014.6943274.

Y. Yuan and M. Q. Meng. *Automatic bleeding frame detection in the wireless capsule endoscopy images*, volume 2015-June. Institute of Electrical and Electronics Engineers Inc., June 2015. doi: 10.1109/ICRA.2015.7139360.

Y. Yuan and M. Q. Meng. Deep learning for polyp recognition in wireless capsule endoscopy images:. *Medical Physics*, 44(4):1379–1389, apr 2017. ISSN 00942405. doi: 10.1002/mp.12147.

Y. Yuan, D. Li, and M. Q.-H. Meng. Automatic polyp detection via a novel unified bottom-up and top-down saliency approach. *IEEE journal of biomedical and health informatics*, 22(4):1250–1260, 2017a.

Y. Yuan, W. Qin, B. Ibragimov, B. Han, and L. Xing. *RIIS-DenseNet: Rotation-Invariant and Image Similarity Constrained Densely Connected Convolutional Network for Polyp Detection*. Springer, 09 2018.

Y. Yuan, W. Qin, B. Ibragimov, G. Zhang, B. Han, M. Q. Meng, and L. Xing. Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition. *IEEE Transactions on Automation Science and Engineering*, 17: 574–583, 4 2020. ISSN 15583783. doi: 10.1109/TASE.2019.2936645.

Z. Yuan, M. Izadyyazdanabadi, D. Mokkapati, R. Panvalkar, J. Shin, N. Tajbakhsh, S. Gurudu, and J. Liang. *Automatic polyp detection in colonoscopy videos*, volume 10133. SPIE, 2017b. doi: 10.1117/12.2254671.

D. E. Yung, E. Rondonotti, and A. Koulaouzidis. Capsule colonoscopy — a concise clinical overview of current status. *Annals of translational medicine*, 4(20), 2016.

S. M. Zaman, M. M. Hasan, R. I. Sakline, D. Das, and M. A. Alam. *A Comparative Analysis of Optimizers in Recurrent Neural Networks for Text Classification*. Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/CSDE53843.2021.9718394.

J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):1–17, 2018. ISSN 15491676. doi: 10.1371/journal.pmed.1002683.

H. Zhang, H. Song, S. Li, M. Zhou, and D. Song. A survey of controllable text generation using transformer-based pre-trained language models, 2022.

R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. W. Lau, and C. C. Y. Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE Journal of Biomedical and Health Informatics*, 21(1):41–47, Jan 2017. ISSN 2168-2194. doi: 10.1109/JBHI.2016.2635662.

X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, and J. Si. Real-time gastric polyp detection using convolutional neural networks. *PLOS ONE*, 14 (3):1–16, 03 2019. doi: 10.1371/journal.pone.0214133.

Q. Zhao and M. Q. . Meng. *Polyp detection in wireless capsule endoscopy images using novel color texture features*. Institute of Electrical and Electronics Engineers Inc., June 2011. doi: 10.1109/WCICA.2011.5970656.

Q. Zhao, G. E. Mullin, M. Q.-H. Meng, T. Dassopoulos, and R. Kumar. A general framework for wireless capsule endoscopy study synopsis. *Computerized Medical Imaging and Graphics*, 41:108 – 116, 2015. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2014.05.011. Machine Learning in Medical Imaging.

S. Zhao, B. Li, C. Reed, P. Xu, and K. Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv*, 2 2020. doi: 10.48550/arxiv.2002.12169.

X. Zhao, C. Fang, F. Gao, D.-j. Fan, X. Lin, and G. Li. Deep Transformers for fast small intestine grounding in Capsule Endoscope Video. VIDEO School of Data and Computer

Science , Sun Yat-Sen University , Guangzhou , China School of Artifical Intelligence , Xidian University , Xi ' an , China The Sixth Affiliated Ho. *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 150–154, 2021.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. *Learning Deep Features for Discriminative Localization*. IEEE Computer Society, Los Alamitos, CA, USA, jun 2016. doi: 10.1109/CVPR.2016.319.

C. Zhou, K. Qiu, C. Chen, D. Zhang, and Y. Guo. Video super-resolution for wireless capsule endoscopy imaging sensor. *IEEE Sensors Journal*, 22(17):17283–17290, 2022.

R. Zhu, R. Zhang, and D. Xue. *Lesion detection of endoscopy images based on convolutional neural network features*. Institute of Electrical and Electronics Engineers Inc., Oct 2015. doi: 10.1109/CISP.2015.7407907.

W. Zhu, L. Xie, J. Han, and X. Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12, 2020.

Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. J. Deng. *Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network*. Institute of Electrical and Electronics Engineers Inc., July 2015. doi: 10.1109/ICDSP.2015.7252086.

L. L. Zwinger, B. Siegmund, A. Stroux, A. Adler, W. Veltzke-Schlieker, R. Wentrup, C. Jürgensen, B. Wiedenmann, F. Wiedbrauck, S. Hollerbach, et al. Capsocam sv-1 versus pillcam sb 3 in the detection of obscure gastrointestinal bleeding. *Journal of clinical gastroenterology*, 53(3):e101–e106, 2019.